

Enzyklopädie der Psychologie

ENZYKLOPÄDIE DER PSYCHOLOGIE

In Verbindung mit der
Deutschen Gesellschaft für Psychologie

herausgegeben von

Prof. Dr. Carl F. Graumann, Heidelberg

Prof. Dr. Theo Herrmann, Mannheim

Prof. Dr. Hans Hörmann, Bochum

Prof. Dr. Martin Irle, Mannheim

Prof. Dr. Dr. h.c. Hans Thomae, Bonn

Prof. Dr. Franz E. Weinert, München

Themenbereich B

Methodologie und Methoden

Serie I

Forschungsmethoden der Psychologie

Band 2

Datenerhebung



Verlag für Psychologie · Dr. C. J. Hogrefe
Göttingen · Toronto · Zürich

Datenerhebung

Herausgegeben von

Prof. Dr. Hubert Feger, Hamburg
und Prof. Dr. Jürgen Bredenkamp, Trier



Verlag für Psychologie · Dr. C. J. Hogrefe
Göttingen · Toronto · Zürich

© by Verlag für Psychologie Dr. C. J. Hogrefe, Göttingen 1983
Alle Rechte, insbesondere das der Übersetzung in fremde Sprachen, vorbehalten.

Gesamtherstellung: Allgäuer Zeitungsverlag GmbH, 8960 Kempten (Allgäu)
Printed in Germany

ISBN 3 8017 0512 9

Autorenverzeichnis

Prof. Dr. Hubert Feger

Psychologisches Institut I
der Universität Hamburg

II. Obergeschoß
von-Melle-Park 6
D - 2000 Hamburg 13

Prof. Dr. Bernd Schäfer

Fachbereich 21 - Psychologie der
Westf.-Wilhelms-Universität Münster

Fliednerstraße 21
D - 4400 Münster

Prof. Dr. Carl F. Graumann

Psychologisches Institut
der Universität Heidelberg

Hauptstraße 47-51
D - 6900 Heidelberg

Prof. Dr. Ralf Schwarzer

Institut für Psychologie
der Freien Universität Berlin

Habelschwerdter-Allee 45
D - 1000 Berlin 33

Dipl.-Psych. Manfred Heinz

Fachbereich I
Abteilung Pädagogik
der Universität Trier

Tarforst
D - 5500 Trier

Prof. Dr. Dr. h.c. Hans Thomae

Psychologisches Institut
der Universität Bonn

An der Schloßkirche 1
D - 5300 Bonn

Prof. Dr. Wolf-R. Minsel

Lindenstraße 10
D - 5501 Ralingen/Wintersdorf

Dr. Ulrich Tränkle

Psychologisches Institut
der Universität Münster
Fachbereich Psychologie

Schlaunstraße 2
D - 4400 Münster

Priv.-Doz. Dr. Franz Petermann

Psychologisches Institut
der Universität Bonn

An der Schloßkirche 1
D - 5300 Bonn

Prof. Dr. Udo Undeutsch

Psychologisches Institut I
der Universität Köln

Haedenkampstraße 2
D - 5000 Köln 41

Vorwort

Angeregt von Hans Thomae und Gustav Adolf Lienert haben sich die Herausgeber die Aufgabe gestellt, ein Handbuch der Allgemeinen Psychologischen Methodenlehre zu edieren. Ein solcher Band war schon bei der ersten Konzipierung der Handbuchreihe geplant, scheiterte zunächst aber an Schwierigkeiten, zu denen auch der Stand dieser psychologischen Disziplin in den fünfziger und sechziger Jahren im deutschsprachigen Raum gehörte. Auch für die nun vorliegende Ausgabe glaubten wir, nicht ganz auf Unterstützung aus dem nicht deutschsprachigen Raum verzichten zu sollen. Da nun im Rahmen des Handbuches der Psychologie der Methodenteil nicht mehr erscheinen konnte, wurde er auf Wunsch des Verlages Bestandteil der Enzyklopädie.

Aus der Entstehungsgeschichte heraus und aus der Tatsache, daß es eine vergleichbare Publikation auch im Angelsächsischen nicht gibt, wird verständlich, daß diese Bände im wesentlichen zwei Funktionen erfüllen möchten: eine systematische Darstellung des gegenwärtigen Standes der psychologischen Methodenlehre zu geben und einige jener Lücken zu füllen, die sich aus verschiedenartigen Gründen bei der Darstellung der Methoden in den früheren Handbuchbänden bisher ergeben hatten. In einigen Handbuchbänden ist die für den jeweiligen Bereich spezifische Methodenlehre dargestellt worden, beispielsweise von Thomae (1959) die der Entwicklungspsychologie und von Graumann (1965) die der Motivationsforschung. Etwa verbliebene Lücken jener speziellen Methodenlehren wird eine Allgemeine Methodenlehre nicht füllen wollen. Einige der in der früheren Handbuchreihe erschienenen Arbeiten, besonders solche im sozialpsychologischen Doppelband (z.B. von Crano & Frenz, 1969, sowie Bredenkamp, 1969) sind jedoch in allen Bereichen der Psychologie von Bedeutung und in diesem Sinn Beiträge zu einer Allgemeinen Psychologischen Methodenlehre.

Die Annahme, daß es sinnvoll sei, von einer Allgemeinen Psychologischen Methodenlehre zu sprechen, hat sich in unserem Fach erst im Laufe der Zeit durchgesetzt; erst die 1972 verabschiedete Rahmenprüfungsordnung sieht ein Fach „Methodenlehre“ vor - dann allerdings an erster Stelle im Kanon der Fächer. Von Allgemeiner Methodenlehre zu sprechen, heißt davon auszugehen, daß es einen genügend großen und tragfähigen Bestand von Prinzipien und Verfahrensregeln gibt, der grundsätzlich in allen Bereichen der Psycholo-

gie und im wesentlichen in gleicher Weise anwendbar ist. Wir beschränken uns auf Forschungsmethoden, schließen also z.B. therapeutische Methoden, überhaupt - wegen ihres anderen Zweckes - alle Interventionsmethoden aus. Wir fassen Psychologie als eine empirische Wissenschaft auf und beschränken uns daher auf Verfahren der Erhebung und Auswertung von Beobachtungen, gehen also beispielsweise auf hermeneutische oder rein mathematische und logische Methoden nicht ein.

Um die Gliederung zu verdeutlichen und die Auswahl der Themen zu begründen, aber auch um eine Lesehilfe zu geben, sei ein idealisiertes Konzept des empirischen Forschungsprozesses skizziert. Bei der Konzeption dieses Modells sind wir davon ausgegangen, daß zu empirischen Hypothesen Beobachtungen angestellt werden, die mit den prognostizierten Beobachtungsergebnissen verglichen werden. Unter empirischen Hypothesen werden Aussagen über den Zusammenhang von wenigstens zwei Variablen verstanden, die aufgrund des Ausgangs des Vergleichs beibehalten oder abgeändert werden. Verglichen werden erhaltene mit prognostizierten Daten, denen ein Datenmodell etwa in Form einer axiomatisierten Meßtheorie zugrunde liegt. Weichen die prognost-

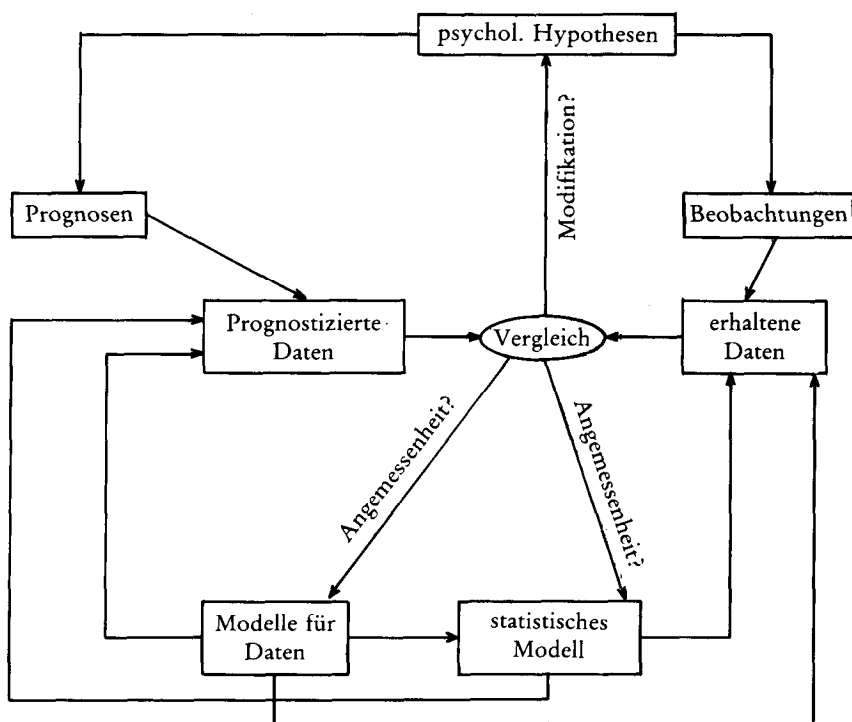


Abbildung 1: Idealisiertes Modell des Forschungsprozesses

stizierten Daten von den erhaltenen ab und ist ein solches Ergebnis reproduzierbar, so kann man sich zu einer Abänderung der empirischen Hypothese entschließen. Führt der Vergleich zu einer Übereinstimmung zwischen beiden Datenstrukturen, wird die empirische Hypothese beibehalten.

Abbildung 1 verdeutlicht, daß dieser Vergleich nicht für schlicht theoriefrei beobachtbare Gegebenheiten durchgeführt wird: Hinter den Daten stehen eine Datentheorie und ein statistisches Strukturmodell. Nach einer Datentheorie müssen bestimmte Voraussetzungen erfüllt sein, damit etwa die empirischen Relationen auf einer Intervallskala in ein numerisches Relativ abgebildet werden. Der Vergleich zwischen prognostizierten und erhaltenen Daten auf dieser Ebene führt zu einer Aussage über die Angemessenheit der Theorie für die Daten. So kann sich etwa ergeben, daß die Daten „nur“ das Niveau einer Ordinalskala erreichen. Dies kann Auswirkungen für die Formulierung des statistischen Strukturmodells haben. Z.B. kann man sich entschließen, mittels sog. nonparametrischer Verfahren für Rangskalen die empirische Hypothese zu prüfen. Zu einem derartigen Entschluß kann man auch gelangen, wenn die Annahmen des statistischen Strukturmodells sich als unzutreffend erweisen. So kann man sich zu einer nonparametrischen statistischen Analyse entschließen, wenn die Daten etwa nicht die Voraussetzung einer Normalverteilung erfüllen. Statistisches Strukturmodell und Datentheorie sind, im Gegensatz zu der empirischen Hypothese, nicht aufgrund von Daten zu modifizieren. Sie können nur in der jeweiligen Situation angemessen oder unangemessen sein.

Der Vergleich zwischen prognostizierten und erhaltenen Daten geschieht auf drei Ebenen. Es sollte deutlich sein, daß auf jeder Ebene andere Prognosen entstehen, die in Abb. 1 zusammengefaßt sind. Wichtig ist, daß eine Bewertung der empirischen Hypothese nicht mittels theoriefrei angestellten Beobachtungen erfolgt, sondern daß hinter den Daten andere Theorien stehen, die für eine Untersuchungssituation angemessen oder nicht angemessen sein können.

Die Komponenten der Abbildung 1 seien anhand eines Beispiels aus der Wahrnehmungspsychologie verdeutlicht. Die empirische Hypothese, die geprüft werden soll, sei das Fechnersche Gesetz (vgl. dazu Tack in Band 3). Die Beobachtungssituation wird derart festgelegt, daß bestimmte Gewichte, in zufälliger Reihenfolge dargeboten, auf einer siebenstufigen Kategorienskala hinsichtlich ihrer Schwere beurteilt werden sollen. Abgesehen wird von weiteren Beobachtungshinsichten (z.B. vom Ausmaß der Schweißsekretion beim Anheben der Gewichte). Erwartet wird, daß das kategoriale Urteil logarithmisch vom Gewicht abhängt. Mit dieser Erwartung werden die tatsächlich angefallenen Daten statistisch verglichen. Dieser statistische Test zur Prüfung des Fechnerschen Gesetzes unterliegt bestimmten Verteilungsannahmen, die getestet werden können. So muß etwa angenommen werden, daß die Wechselwirkungseffekte „log Gewicht x Versuchspersonen“ normalverteilt sind. Mit dieser „Prognose“ kann die empirische Verteilung der Wechselwirkungseffekte verglichen und damit die Angemessenheit des statistischen Verfahrens ge-

prüft werden. Ferner muß unterstellt werden, daß die Kategorienskala eine Intervallskala der subjektiven Empfindungsgrößen ist. Diese Annahme kann geprüft werden, indem die Axiome endlicher absoluter Differenzenstrukturen daraufhin getestet werden, ob sie erfüllt sind (vgl. Westermann 1980). Ist dies nicht der Fall, so sind die Axiome dieser Meßstruktur damit nicht falsifiziert, sondern sie sind für die Untersuchung nicht erfüllt. In diesem Fall kann dann auch die Kategorienskala zur Prüfung des Fechnerschen Gesetzes nicht angewendet werden, und es ist nach einem anderen Skalierverfahren zu suchen, für das die Axiome endlicher absoluter Differenzenstrukturen erfüllt sind. Das Ausweichen auf nonparametrische Tests zur Prüfung des Fechnerschen Gesetzes, wenn die subjektiven Größen keine Intervallskala konstituieren, ist nicht möglich, da mit der Ableitung des Fechnerschen Gesetzes die Annahme einer Intervallskala verknüpft ist (vgl. Luce 1959, 1962; Rozeboom 1962).

Die vorliegenden Bände gliedern sich, anhand der Abb. 1 erläutert, wie folgt: Gegenstand des ersten Teiles sind die wissenschaftlichen Beobachtungen; auch Aussagen über die Selbstbeobachtung. Gemäß Abb. 1 handelt es sich um die „Basis“ für die Beurteilung von empirischen Hypothesen. Hierzu sind auch die Artikel des Bandes 2 zu zählen, die speziellen Erhebungsmethoden wie dem Q-Sort, dem semantischen Differential etc. gewidmet sind. Es geht um Forschungsmethoden zur Erhebung von Daten zur Prüfung spezieller Hypothesen; Erhebungsmethoden wie diese ergaben sich aus dem Bemühen, das Gewinnen von Beobachtungen den jeweils untersuchten Gegebenheiten spezifisch anzupassen. In Band 3 wird das, was in Abb. 1 „Modelle für Daten“ genannt wird, behandelt, wobei zwischen „Messung und Skalierung“ und „Tests“ unterschieden wird. Band 4 ist dem Bereich gewidmet, der in Abb. 1 „statistisches Modell“ heißt, wobei allerdings nonparametrische Prüfverfahren unberücksichtigt bleiben, über die ein umfangreiches Handbuch von Lienert (1973, 1975, 1978) in deutscher Sprache orientiert. Schließlich finden sich im Band 5 Artikel, die über die Modellierung psychischer Prozesse und die Prüfung empirischer Hypothesen unterrichten.

Wie bereits ausgeführt, gibt Abb. 1 ein idealisiertes Modell des Forschungsprozesses wieder, das uns bei der Ordnung der Teile dieser Bände leitete. Selbstverständlich wurde jeder Beitrag „für sich“ geschrieben, ohne daß die Autoren auf diese Leitvorstellungen verpflichtet wurden. Dies sollte bei einer Beurteilung der Abb. 1 im Hinblick auf die vorliegenden Artikel berücksichtigt werden. Vermutlich wären auch andere Ordnungsschemata denkbar gewesen, die zu teilweise anderen Zusammenstellungen der Artikel und/oder Selektionen der zu behandelnden Themen geführt hätten. Wichtig ist vor allem, daß jeder Artikel für sich genommen dem Forscher die Informationen und Literaturhinweise gibt, die er sich für seine Arbeit erhofft.

Die Planung dieser Bände sah die Berücksichtigung weiterer Artikel vor, die nicht erschienen sind, da die Publikation dieses Bandes sich in unvertretbarem Maße verzögert hätte. Wir verweisen hier vor allem auf die Faktorenanalyse,

die statistischen Kausalanalysen und die nonmetrische Skalierung, auf die teilweise in einigen Beiträgen kurz eingegangen wird, die aber eigenständig repräsentiert sein sollten. Zu diesen Themen sei deshalb auf neuere deutschsprachige Literatur verwiesen. über Faktorenanalyse handeln die beiden Bücher von Revenstorf (1976, 1980) und der umfassende Artikel Pawliks (1977), wichtige Beiträge zu statistischen Kausalanalysen und eine Einführung in dieses Gebiet enthält der dreibändige Reader von Hummell und Ziegler (1976), und zur nonmetrischen wie metrischen insbesondere theorieprüfenden Skalierung verweisen wir auf Borg (1982). Auch die statistischen Einzelfallanalysen sollten in dieser Enzyklopädie vertreten sein. Da aber der Beitrag von Huber in Band VIII des Handbuches der Psychologie gerade fertiggestellt war, verzichteten wir auf eine weitere Publikation zu diesem Thema.

Verschiedene Themen werden in diesen Bänden mehrfach angesprochen, beispielsweise die Raschskalierung in mehreren Artikeln. Allerdings handelt es sich nach Meinung der Herausgeber nicht wirklich um Redundanzen, vielmehr stehen die Ausführungen jeweils in einem anderen Kontext und gewinnen von daher ihre eigene Berechtigung. Deshalb haben die Herausgeber die Autoren nicht gebeten, diese Passagen zu streichen.

Die Herausgeber hatten auch erwogen, alle Autoren auf die gleiche Notation und sogar auf die gleiche Terminologie (aber welche?) festzulegen. Uns schien jedoch der dafür erforderliche Aufwand in keinem Verhältnis zum Gewinn an Lesbarkeit zu stehen, und eine einheitliche Terminologie hätte im günstigsten Fall eine Homogenität der Methodenlehre vorgetäuscht, die nicht vorhanden ist; im ungünstigsten Fall hätte sie die verschiedenen Ansätze verfälscht. Die Anfertigung der Literaturverzeichnisse sollte den Konventionen in der Zeitschrift für Sozialpsychologie folgen. Auch hierauf haben wir letztendlich nicht bestanden; geachtet wurde lediglich auf Vollständigkeit der Angaben.

Die einzelnen Artikel setzen auf unterschiedlichem Niveau Vorkenntnisse voraus. Die meisten gehen von jenem Wissensstand aus, über den ein Psychologiestudent nach gut bestandener Prüfung im Fach Methodenlehre verfügen sollte. Den Charakter einer Einführung haben die Bände also nicht. Neben den zahlreichen englischsprachigen Einführungen können wir auf mehrere deutsche oder übersetzte verweisen (Bartenwerfer & Raatz 1979, Crano & Brewer 1975, Friedrichs 1973, Kerlinger 1978, Klapprott 1975, Selg & Bauer 1971, Traxel 1964, Wottawa 1977). Gesamtdarstellungen, von denen große Teile auch für Psychologen relevant sind, finden sich in den Nachbarfächern, besonders der Soziologie (König 1962, van Koolwijk & Wicken-Mayser 1974). Erwähnt seien auch einige wenige psychologische Zeitschriften, in denen regelmäßig und gehäuft Beiträge zur Methodenlehre erscheinen: *Applied Psychological Measurement*, *British Journal of Mathematical and Statistical Psychology*, *Educational and Psychological Measurement*, *Journal of Mathemati-*

cal Psychology, Multivariate Behavioral Research, Psychometrika, Psychological Bulletin, sowie die regelmäßigen übersichten im Annual Review of Psychology. Im deutschen Sprachbereich gibt es noch keine ausschließlich methodenorientierte psychologische Zeitschrift, jedoch finden sich in allen wissenschaftlichen Zeitschriften unseres Faches regelmäßig Publikationen zu methodischen Problemen der Psychologie.

Es wäre reizvoll, die historische Entwicklung im Detail nachzuzeichnen, die zur Begründung der Allgemeinen Psychologischen Methodenlehre geführt hat. Wie für viele andere Disziplinen unseres Faches waren die Quellen heterogen, und zwischen verschiedenen Strömungen gab es kaum Berührungen. Die älteste Tradition hat mit der Psychophysik die Behandlung der Frage aufzuweisen, ob und wie die Variablen der Psychologie meßbar sind. Unabhängig von der Psychophysik tauchte das Meßproblem im Rahmen der Testtheorie auf, und schließlich, über Thurstone direkt mit der Psychophysik verbunden, in der Einstellungsmessung. Erst die letzten beiden Jahrzehnte haben zu Querverbindungen zwischen diesen Meßtraditionen geführt.

Nicht viel jünger ist die Tradition, mit der ein Student unseres Faches meistens den ersten Kontakt hat, mit der Statistik, insbesondere der Inferenzstatistik, die durchweg mit der Planung und Auswertung von Experimenten verbunden ist. Als erste deutschsprachige Einführungen sind hier die von Lazarsfeld (1929), Mittenecker (1952) und Hofstätter (1953) zu erwähnen. Querverbindungen zwischen Skalierung und Inferenzstatistik beobachtet man ebenfalls in den letzten Jahrzehnten.

Weniger stürmisch als bei den Auswertungsverfahren, die mehr und mehr als formale Modelle des jeweils untersuchten Bereiches begriffen werden, entwickelten sich die Erhebungsverfahren, die oft als Sammlung von Erfahrungen und praktischen Ratschlägen erscheinen. Guttman's (1959) Facettentheorie und die Analyse von Erhebungsverfahren durch Coombs (1964, insb. Kap. 2) könnten erste Ansätze für eine allgemeine Theorie von Erhebungsverfahren darstellen. Eine Geschichte der psychologischen Methodenlehre ist allerdings erst noch zu schreiben.

Die Herausgeber haben einer Reihe von Personen für ihre Mitarbeit zu danken. Die Diplom-Psychologen E. Erdfelder, J. Funke, T. Kindermann und P. Mann sowie die Studenten E. Hafner, M. Kruppert, M. Meyer und A. Schrameier (alle Trier) haben an der Erstellung der Register mitgewirkt und die Korrekturen besorgt. In Hamburg haben Dipl.-Psych. U. Droge und Dipl.-Psych. F. Mohazab diese Arbeiten übernommen. Gedankt sei auch den Autoren, die ihre Beiträge fristgerecht fertiggestellt haben, für ihre Geduld. Die Artikel sind z. T. 1980, einige im Frühjahr 1981 bei uns eingetroffen.

J. B.
H. F.

Literatur

- Bartenwerfer, H. & Raatz, U. Einführung in die Psychologie, Band 6: Methoden der Psychologie. Wiesbaden: Akademische Verlagsgesellschaft & Bern: Huber, 1979.
- Borg, I. Angewandte multidimensionale Skalierungen. Berlin: Springer, 1982.
- Bredenkamp, J. Experiment und Feldexperiment. In: Graumann, C. F. (Hrsg.): Handbuch der Psychologie, Band 7/I, Sozialpsychologie. Göttingen: Hogrefe, 1969, S. 332-374.
- Coombs, C. H. A theory of data. New York: Wiley, 1964.
- v. Cranach, M. & Frenz, H. G. Systematische Beobachtung. In: Graumann, C. F. (Hrsg.): Handbuch der Psychologie, Band 7/I, Sozialpsychologie. Göttingen: Hogrefe, 1969, S. 269-331.
- Crano, W. D. & Brewer, M. B. Einführung in die sozialpsychologische Forschung. Köln: Kiepenheuer & Witsch, 1975. (Original: Principles of research in social psychology. New York: McGraw-Hill, 1973).
- Friedrichs, J. Methoden empirischer Sozialforschung. Reinbek b. Hamburg: Rowohlt Taschenbuch, 1973.
- Graumann, C. F. Methoden der Motivationsforschung. In: Thomae, H. (Hrsg.): Handbuch der Psychologie, Band 2, Allgemeine Psychologie, II. Motivation. Göttingen: Hogrefe, 1965, S. 123-202.
- Guttman, L. A structural theory of intergroup beliefs and action, American Sociological Review 1959, 24, 318-328.
- Hofstätter, P. R. Einführung in die quantitativen Methoden der Psychologie. München: Barth, 1953.
- Hummell, H. J. & Ziegler, R. (Hrsg.). Korrelation und Kausalität. 3 Bände. Stuttgart: Enke, 1976.
- Kerlinger, F. N. Grundlagen der Sozialwissenschaften. Weinheim: Beltz, 1978. (Original: Foundations of behavioral research. New York: Holt, Rinehart & Winston 1964, 1973).
- Klapprott, J. Einführung in die psychologische Methodik. Stuttgart: Kohlhammer (Urban Taschenbücher, Band 214), 1975.
- König, R. (Hrsg.). Handbuch der empirischen Sozialforschung. Stuttgart: Enke, 1962.
- van Koolwijk, J. & Wieken-Mayser, M. (Hrsg.). Techniken der empirischen Sozialforschung. 8 Bände. München: Oldenbourg, 1974f.
- Lazarsfeld, P. F. Statistisches Praktikum für Psychologen und Lehrer. Jena: Fischer, 1929.
- Lienert, G. A. Verteilungsfreie Methoden in der Biostatistik. 3 Bände. Meisenheim am Glan: Anton Hain, 1973, 1975, 1978.
- Luce, R. D. On the possible psychophysical laws. Psychological Review 1959, 66, 81-95.

- Luce, R. D. Comments on Rozeboom's criticisms of "On the possible psychophysical laws". *Psychological Review* 1962, 69, 548-551.
- Mittenecker, E. Planung und statistische Auswertung von Experimenten. Wien: Deuticke, 1952.
- Pawlik, K. Faktorenanalytische Persönlichkeitsforschung. In: Strube, G. (Hrsg.): *Die Psychologie des 20. Jahrhunderts. Kindler Enzyklopädie, Band 5.* Zürich: Kindler, 1977, S. 617-712.
- Revenstorf, D. *Lehrbuch der Faktorenanalyse*, Stuttgart: Kohlhammer, 1976.
- Revenstorf, D. *Faktorenanalyse*. Stuttgart: Kohlhammer, 1980.
- Rozeboom, W. W. The untenability of Luce's principle. *Psychological Review* 1962, 69, 542-547.
- Rozeboom, W. W. Comment. *Psychological Review* 1962, 69, 552.
- Selg, H. & Bauer, W. *Forschungsmethoden der Psychologie*. Stuttgart: Kohlhammer (Urban Taschenbücher, Band 121), 1971.
- Thomae, H. *Forschungsmethoden der Entwicklungspsychologie*. In: Thomae, H. (Hrsg.): *Handbuch der Psychologie, Band 3, Entwicklungspsychologie*. Göttingen: Hogrefe, 1959, S. 46-75.
- Traxel, W. *Einführung in die Methodik der Psychologie*. Bern: Huber, 1964.
- Westermann, R. Die empirische Überprüfung des Niveaus psychologischer Skalen. *Zeitschrift für Psychologie* 1980, 188, 45-68.
- Wottawa, H. *Psychologische Methodenlehre*. München: Juventa, 1977.

Inhaltsverzeichnis

1. Kapitel: Planung und Bewertung von wissenschaftlichen Beobachtungen. Von Hubert Feger

1. Übersicht und Systematik	1
2. Arten von Beobachtungen	3
2.1 Allgemeine Übersicht	3
2.2 Teilnehmende Beobachtung	5
3. Die Planung von Beobachtungen	6
3.1 Das <i>Universum von Beobachtungen</i>	6
3.2 <i>Bestimmen der Beobachtungseinheit</i>	10
3.3 <i>Kategoriensysteme</i>	12
3.4 <i>Auswahlen aus dem Universum der Beobachtungen</i>	15
3.4.1 Auswahl von Personen	15
3.4.2 Auswahl und Schulung von Beobachtern	17
3.4.3 Auswahl des zu beobachtenden Verhaltens	18
3.4.4 übergreifende Auswahlstrategien	20
4. Die Bewertung von Beobachtungen	22
5. Die Reproduzierbarkeit von Beobachtungen	23
5.1 <i>Übereinstimmungsmaße für nominalskalierte Daten</i>	26
5.1.1 Prozentuale Übereinstimmung und allgemeine Vorüberlegungen	26
5.1.2 Systematik einiger Übereinstimmungsmaße für nominalskalierte Daten	29
5.2 <i>Übereinstimmungsmaße für ordinalskalierte Daten</i>	35
5.3 <i>Übereinstimmungsmaße für intervallskalierte Daten</i>	37
5.3.1 Einfache varianzanalytische Ansätze und Intraklassen-Koeffizienten	37
5.3.2 Generalisierbarkeitsstudien	41
5.3.3 Pfadanalytische Modelle für die Reliabilitätsprüfung	42
5.4 <i>Besondere Erhebungspläne</i>	44

5.5 Die Berücksichtigung von Reliabilitätskenntnissen bei der weiteren Datenauswertung	46
6. Validität von Beobachtungen	48
6.1 Konstruktvalidierung	50
6.2 Neuere Entwicklungen zur Analyse von multitrait-multimethod Matrizen	54

2. Kapitel: Beobachtung und Beschreibung von Erleben und Verhalten. Von Hubert Feger und Carl F. Graumann

1. Vorbemerkungen zu Thema und Terminologie	76
2. Formen der Erlebnisbeschreibung	77
2.1 Selbstbeobachtung und Erlebnisbeschreibung als Methoden und Themen der Psychologie	77
2.2 Selbstbeobachtung und Experiment: Die Begründung der wissenschaftlichen Psychologie	80
2.3 Die systematische experimentelle Selbstbeobachtung	84
2.3.1 Die konkrete Vorgehensweise	84
2.3.2 Maßnahmen zur Sicherung der Ergebnisse	85
2.3.3 Begründung der Möglichkeit von Selbstbeobachtung	87
2.3.4 Anmerkungen zu typischen Ergebnissen	88
2.4 Die behavioristische Kritik der „Introspektion“	89
2.5 Die Technik des lauten Denkens	91
2.6 Phänomendeskription	91
2.7 Behavioristische Selbstwahrnehmung	93
2.8 Neuere Untersuchungen über bildhafte Vorstellungen	94
2.9 Methoden der Metakognitionsforschung	97
3. Aktuelle Probleme der Verhaltensbeobachtung	99
3.1 Der Gegenstandspsychologischer Verhaltensbeobachtung	99
3.2 Analyse des Beobachters als Meßinstrument	101
3.2.1 Die Ermittlung von „Fehlern“	101
3.2.2 Der Einfluß von semantischen Gedächtnisstrukturen auf Verhaltensbeschreibungen	102
3.2.3 Die Theorie der Signalentdeckung: Der Beobachter als Sensorium und als Entscheidungsinstanz	107
3.2.4 Verhaltenseinschätzungen als Testscores	109
3.2.5 Brunswiks probabilistischer Funktionalismus: Beobachtung als Leistung	110
3.3 Die Wahl von Beobachtungseinheiten durch Beobachter.	112
3.4 Der Entstehungsprozeß von Beschreibungen	114
3.5 Verhaltenseinschätzung (behavioral assessment)	116
3.5.1 Die Verlässlichkeit von Selbstberichten und Fremdbeobachtungen	118
3.5.2 Reaktivität	120
3.5.3 Einflüsse bestehender Erwartungen der Beobachter	123

3. Kapitel: Das Q-Sort-Verfahren.

Von Wolf-Rüdiger Minsel und Manfred Heinz

1. Zur Einordnung des Q-Sort-Verfahrens	135
2. Beispiel eines Q-Sort-Verfahrens	136
3. Anwendung des Q-Sort-Verfahrens	140
4. Probleme des Q-Sort-Verfahrens	141
4.1 <i>Itemselektion und Itemorganisation</i>	141
4.2 <i>Verteilungsform</i>	143
4.3 <i>Auswertung</i>	144
4.4 <i>Gütekriterien</i>	144
4.4.1 <i>Reliabilität</i>	144
4.4.2 <i>Validität</i>	145
4.5 <i>Qualität der Daten</i>	147
5. Bedeutung des Q-Sort-Verfahrens	148

4. Kapitel: Semantische Differential Technik.

Von Bernd Schäfer

1. Einleitung	154
1.1 <i>Zugrundeliegende Modelle</i>	154
1.1.1 <i>Verhaltensmodell (representational mediation theory)</i>	155
1.1.2 <i>Meßmodell</i>	156
1.1.3 <i>Raummodell</i>	156
1.2 <i>Integration der Modelle</i>	157
2. Ordnung von SD-Daten: Architektur eines universellen Bedeutungsraumes	158
2.1 <i>Skalen-Kovariation: Generalität der EPA-Struktur</i>	159
2.1.1 <i>Grundlegende Befunde (The Measurement of Meaning: Osgood et al. 1957)</i>	159
2.1.2 <i>Berücksichtigung der verfügbaren Varianz von SD-Daten</i>	161
Daten-Reduktionstechniken	161
Konzeptvarianz	162
EXKURS: Affektive (konnotative) und denotative Bedeutung	164
2.1.3 <i>Variationen des Modus der Dimensionsanalyse</i>	167
2.1.4 <i>Transkulturelle Stabilität</i>	168
2.1.5 <i>Interindividuelle Unterschiede</i>	169
2.2 <i>Interaktionsvarianz: Konzept-Skalen-Interaktion</i>	172
2.3 <i>„Fehlervarianz“</i>	178
2.3.1 <i>Systematische Urteilsfehler</i>	178

Extremisierung	178
Soziale Erwünschtheit	179
2.3.2 Zufallsfehler-Reliabilität von SD-Urteilen	181
3. Metrische Eigenschaften von SD-Skalen: ‚Statik‘ des semantischen Raumes	184
3.1 <i>Bipolarität</i>	184
3.2 <i>Intervallgleichheit</i>	187
3.3 <i>Nullpunktlage</i>	188
4. Wahl von SD-Skalen zur Exploration von Bedeutungs-Räumen: Konstruktion von Semantischen Differentials	189
4.1 <i>Merkmals-Relevanz</i>	191
4.2 <i>Merkmals-Polarität</i>	194
4.3 <i>Dimensionale Repräsentativität</i>	195
4.4 <i>Variationen der Präsentationsweise</i>	196
4.4.1 Reihenfolge der Konzept-Skalenkombination	196
4.4.2 Verankerung der Skalen	197
4.4.3 Zahl der Antwortkategorien	198
4.5 <i>Varianten der Technik</i>	199

5. Kapitel: Fragebogenkonstruktion. Von Ulrich Tränkle

1. Einführung	222
1.1 <i>Versuch einer Systematik von Fragebogen</i>	222
1.1.1 Einteilungsgesichtspunkte für Fragebogen	222
1.1.2 Grundkonzeptionen von Fragebogen	224
1.1.3 Hauptanwendungsgebiete für Fragebogen	227
1.2 <i>Ansätze zu einer Theorie des Beantwortungsprozesses</i>	229
1.2.1 Determinanten des Antwortverhaltens	229
1.2.2 Antwortgenese	231
1.2.3 Die Frage als Suchbegriff	236
1.3 <i>Einordnung der Fragebogenkonstruktion in die Stadien einer Befragung</i>	238
2. Fragentypen	241
2.1 <i>Zielsetzungen von Fragen</i>	241
2.2 <i>Frageninhalte</i>	243
2.3 <i>Direktheit einer Frage</i>	244
2.4 <i>Formale Fragenkonstruktion</i>	246
2.4.1 Offene und geschlossene Fragen	246
2.4.2 Arten geschlossener Fragen	248
2.4.3 Sonderformen	250
3. Fragenformulierung	251
3.1 <i>Die inhaltliche Konzeption einer Frage</i>	252
3.1.1 Vorüberlegungen	252
3.1.2 Definition des Gegenstandes und Explikation eines Bezugsrahmens	253

3.1.3 Festlegung der Antwortkategorien	254
3.1.4 Verzerrte Fragen	256
3.1.5 Uninformiertheit, Meinungslosigkeit und Urteilsausgewogenheit	259
3.1.6 Antworttendenzen und vorschnelle Antworten	260
3.2 Sprachliche Formulierung der Frage	261
3.2.1 Kriterien für die sprachliche Formulierung	261
3.2.2 Anforderungen an die sprachliche Formulierung	263
3.3 Spezielle Gesichtspunkte der Formulierung von Items für diagnostische Fragebogen	265
3.4 Die Kontrolle von Formulierungseinflüssen	266
4. Reihenfolge der Fragen und Umfang des Fragebogens	267
4.1 Ziele beim Aufbau eines Fragebogens	267
4.2 Motivation der Befragten und Steigerung der Antwortfähigkeit	269
4.3 Reihenfolgeeffekte	270
4.3.1 Kontexteffekte	270
4.3.2 Positionseffekte	273
4.4 Unangenehme und heikle Fragen	274
4.5 Fragen zur Person	275
4.6 Filterfragen und Verzweigungsfragen	275
4.7 Spezielle Gesichtspunkte für die Itemreihenfolge diagnostischer Fragebogen	276
4.8 Überlegungen zur Vermeidung unerwünschter Reihenfolgeeffekte	277
4.9 Fragebogenumfang	278
5. Äußere Gestaltung (Layout) des Fragebogens	279
6. Weitere Aspekte für die Konstruktion von Fragebogen	283
6.1 Anonymität des Befragten und Vertraulichkeit der Antworten	283
6.2 Spezielle Probleme bei unpersönlich-schriftlichen Befragungen	285
6.3 Erprobung und Überarbeitung des Fragebogenentwurfs	287
7. Zukünftige Entwicklung im Bereich der Fragebogenkonstruktion	289

6. Kapitel: Befragung. Von Ralf Schwarzer

1. Begriffsklärung und Übersicht	302
2. Formen und Probleme der Befragung	305
2.1 Schriftliche Befragung	305
2.1.1 Vor- und Nachteile	305
2.1.2 Weitere Probleme und Besonderheiten	306
2.2 Die mündliche Befragung	308
2.2.1 Vor- und Nachteile	308
2.2.2 Der Interviewer	310
2.2.3 Der Befragte	311

2.3 Einige Sonderformen	313
2.3.1 Realkontakt-Befragung	313
2.3.2 Telefoninterview	314
2.3.3 Kinderinterview	315
3. Befragung im Handlungskontext	316
3.1 Befragung und Introspektion	316
3.2 Intendierte Veränderungen im Forschungsprozeß	317

7. Kapitel: Exploration. Von Udo Undeutsch

1. Begriffsbestimmung	321
2. Geschichte	323
3. Qualitative Charakterisierung	325
4. Methodische Prinzipien des explorativen Gesprächs	329
5. Auswertung.	334
6. Leistungsfähigkeit der explorativen Methoden	336
6.1 Reliabilität	339
6.2 Validität	345
6.2.1 Die Validität der Datenerhebung	345
6.2.2 Validität der Bewertung	347
6.2.3 Validität der diagnostischen Verwertung	348

8. Kapitel: Biographische Methode und Einzelfallanalyse. Von Hans Thomae und Franz Petermann

1. Einführung	362
2. Idiographische Persönlichkeitspsychologie und biographische Methode	364
3. Entwicklungspsychologie und humanistische Psychologie	365
4. Probleme psychoanalytischer Biographie	368
5. Biographische Methode als Instrument der Sozialisationsforschung	371
6. Psychologische Streßforschung und biographische Methode	373
7. Das Problem der Objektivität der biographischen Methode	375
8. Vorschläge zur Erhöhung der Objektivität der biographischen Methode	380
9. Das halbstrukturierte Interview und das Problem der Kontrolle der Datengewinnung	383
10. Die Frage der „Einheiten“ der Biographie	384
11. Eine Möglichkeit der statistischen Auswertung von Biographien	386
12. Biographik und Einzelfallanalyse	387
13. Einzelfallanalytische Datensammlung und Versuchsplanung	389
14. Übersicht über statistische Auswertungsmethoden für Einzelfälle	391

Sach-Register	401
Autoren-Register	404

1. Kapitel

Planung und Bewertung von wissenschaftlichen Beobachtungen

*Hubert Feger**

1. Übersicht und Systematik

Alle Aussagen über den jeweiligen Gegenstandsbereich einer empirischen Wissenschaft sind direkt oder indirekt auf Beobachtungen zurückzuführen. In einer (nichtapparativen) Beobachtung stellt sich durch und als unmittelbare Anschauung die Beziehung zwischen Beobachter und Beobachtungsgegenstand her. Beobachtung ist ein Prozeß, in dem Beziehungen zwischen Gegebenheiten hergestellt werden. Welcher Art die zu erfassenden Gegebenheiten und Relationen ihrem Inhalt nach sind, bestimmen die theoretischen und praktischen Fragestellungen; aus dem Gehalt der Theorie wäre abzuleiten, wie Beobachtungen als Beziehungen zwischen Elementen in formaler Hinsicht aufgefaßt werden müssen. Daraus ergeben sich Konsequenzen für die Datenanalyse. Die Prüfung von Aussagen über einen Gegenstandsbereich einer empirischen Wissenschaft wird als Prüfung der Übereinstimmung von aus der Theorie abgeleitetem Sachverhalt mit beobachtetem Sachverhalt angelegt sein, wobei der Prüfung logischer Merkmale der Aussagen selbst, etwa ihre Widerspruchsfreiheit der Status einer notwendigen Voraussetzung zukommt. Letzteres im allgemeinen systematisch zu analysieren ist eine der Aufgaben der Wissenschaftstheorie und der Logik.

Dieses Kapitel befaßt sich nicht mit einer psychologischen Theorie des Beobachtens durch Menschen als Beobachtern, weder mit den Prozessen der Wahrnehmung, Beurteilung und des Gedächtnisses, die Beobachtungen ermöglichen und deren Form und Inhalt beeinflussen, noch mit einer allgemeinen und differentiellen Theorie des Beobachters. Einiges zu dieser Thematik findet sich im Kapitel über Verhaltensbeobachtung und Erlebnisbeschreibung (Feger & Graumann in diesem Band). Auch die Auswertung von Beobachtungen ist

* Dank für wesentliche Hinweise schulde ich C. F. Graumann, I. Borg und K. Westhoff.

im Gegensatz zu vielen Kapiteln nicht Gegenstand dieser Darstellung. Vielmehr schildern und diskutieren wir allgemeine Prinzipien der Planung und Bewertung von Beobachtungen als wissenschaftlichen. Die Planung bezieht sich auf alle Fragen, welche Beobachtungen anzustellen und wie sie durchzuführen seien. Die Bewertung umfaßt alle Fragen nach den Kriterien, die eine Beobachtung als wissenschaftliche erfüllen muß, und wie diese Kriterien zu prüfen sind. Empirische Wissenschaften, und in ihnen verschiedene Forschungsgebiete, unterscheiden sich in dem Ausmaß, in dem der Beobachtungsprozeß standardisiert und instrumentalisiert ist. Der Ersatz des Beobachters durch Apparate beseitigt ganz oder teilweise die Probleme der Subjektivität bei der Auswahl und Bewertung des konkret Beobachteten, nicht jedoch die Probleme der Planung und Bewertung sowie der datentheoretischen Interpretation von Beobachtungen. Nicht behandelt werden deshalb Probleme, die sich durch und bei Gebrauch technischer Apparatur, insbesondere Tonband- und Film- oder Videogeräten ergeben (z. B. Clarke & Ellgring, 1978; Longabaugh, 1980), ebenfalls nicht Fragen der technischen Datenverarbeitung, Ablochung und Speicherung. Zur allgemeinen Methodenlehre gehören definitionsgemäß auch nicht Probleme des Beobachtens, die sich aus der speziellen Natur des Beobachteten ergeben, z.B. Beobachtung von Säuglingen, Tieren in freier Wildbahn etc. Ferner würde die Grenze zur Systematik des Faches überschritten, wenn hier jene theoretischen Ansätze besprochen würden, deren empirische Fundierung zumindest gegenwärtig stark auf Beobachtung angewiesen scheint, wie Ethologie oder Humanökologie. Schließlich haben wir uns bemüht, Überschneidungen mit dem in diesem Band folgenden Kapitel und dem Artikel von v. Cranach und Frenz (1969) zu vermeiden.

Ohne auch nur die klassischen oder jeweils neuesten Arbeiten vollständig auflisten zu können, erwähnen wir vorwiegend deutschsprachige Literatur, die eine Einführung und Übersicht ermöglicht: Peak (1953), König (1962), Graumann (1966), Weick (1968), Jahoda et al. (1968), Grüner (1974), Hutt & Hurt (1974), Faßnacht (1979); insbesondere für die Erfassung nonverbaler Verhaltensweisen Scherer (1974), für verbale Manz (1974), zum ethologischen Ansatz McGrew (1972), für Verhalten in „naturalistic settings“ des Kulturvergleichs: Longabaugh (1980), mit Beispielen aus der pädagogischen Psychologie: Medley & Mitzel (1963), für diese in der Entwicklungspsychologie seit langem gepflegte Methodik: Thomae (1959), Wright (1960), speziell die Kleinkindforschung: Simons & Papousek (1978); mit sozialpsychologischen Schwerpunkten: Heyns & Zander (1953), v. Cranach & Frenz (1969), Duncan & Fiske (1977); aus diagnostischer Perspektive: Hasemann (1964), mit Querverbindung zu Verhaltensmodifikation: Keut & Foster (1977), Mees & Selg (1977); s. auch Feger & Graumann in diesem Band.

2. Arten von Beobachtungen

2.1 Allgemeine Übersicht

Bevor wir auf Planung und Bewertung für Beobachtungen allgemein eingehen, geben wir einen Überblick über verschiedene Arten von Beobachtungen, die in der Psychologie und Nachbarwissenschaften eine Rolle spielen. Sie werden unterschieden nach den Bedingungen, unter denen sie zustande kommen, den Verfahren, wie sie gewonnen werden, und nach Besonderheiten des Beobachtungsgegenstandes.

Graumann (1966, S. 86) akzentuiert, wie sich Beobachtung, auch die noch nicht wissenschaftliche, von Wahrnehmung abhebt: „Die absichtliche, aufmerksam-selektive Art des Wahrnehmens, die ganz bestimmte Aspekte auf Kosten der Bestimmtheit von anderen beachtet, nennen wir Beobachtung. Gegenüber dem üblichen Wahrnehmen ist das beobachtende Verhalten planvoller, selektiver, von einer Suchhaltung bestimmt und von vorneherein auf die Möglichkeit der Auswertung des Beobachteten im Sinne der übergreifenden Absicht gerichtet.“ Wenn die übergreifende Absicht ist, eine wissenschaftliche Annahme zu prüfen, und wenn sie in Planung und Bewertung bestimmten Kriterien genügt, geht die vorwissenschaftliche in die wissenschaftliche Beobachtung über.

Alltägliche und wissenschaftliche Beobachtung unterscheiden sich nicht dadurch, wie Information gewonnen wird, nicht durch die Art der Prozesse, die im Beobachter ablaufen, sondern durch die Ziele, derentwegen die Beobachtungen angestellt werden, und durch die Umstände, die wegen dieser Ziele aufgesucht oder hergestellt werden. Beobachtungen lassen sich nach dem Verfahren, wie sie zustande kommen, und dabei wieder nach verschiedenen Gesichtspunkten klassifizieren. Uns erscheint ein Schema von Graumann (1966) das in Tab. 1 wiedergegeben ist, sehr übersichtlich. Das Schema ist lediglich für die kontrollierte direkte Verhaltensbeobachtung ausgeführt. Soweit möglich sollte man sich diese Klassifikationskriterien nicht dichotom, sondern als Pole *eines* Kontinuums vorstellen. Die erste Klassifikation nach nichtkontrollierten und *kontrollierten* Beobachtungen, d.h. solchen, bei denen die Bedingungen, unter denen sie zustande kamen, bekannt sind, ist für manche Autoren gleichbedeutend mit der Unterscheidung zwischen nichtwissenschaftlicher und wissenschaftlicher Beobachtung, doch sollte nach Graumann gerade diese Klassifikation nicht als Dichotomie angesehen werden. Bezeichnet man eine Beobachtung als direkt, so kann dies in der Literatur dreierlei bedeuten: (1) Zwischen Beobachter und Beobachtetem steht kein Hilfsmittel, kein Apparat, Test o.ä. Graumann spricht dann - als fünftem Kriterium - von *vermittelter* vs. *unvermittelter* Beobachtung. Der Einsatz einer Blickbewegungskamera führt also zu direkten, vermittelten Beobachtungen. (2) Zwischen Beobach-

tung einerseits, Beurteilung sowie Registrierung andererseits liegt kein größerer Zeitraum; die retrospektive Rekonstruktion bei Fallstudien wäre demnach als indirekte Beobachtung zu klassifizieren. Graumann jedoch bezeichnet eine Verhaltensbeobachtung dann als indirekt, wenn (3) sie sich nicht auf das Verhalten selbst, sondern auf dessen Spuren, Auswirkungen und Objektivationen richtet. Zu diesen indirekten Verfahren gehört z.B. die systematische Inhaltsanalyse.

Die *teilnehmende* Beobachtung, in der ein Beobachter für die Beobachteten einen erkennbaren Teil der Beobachtungssituation ausmacht, wird im folgenden Abschnitt behandelt. Als *unwissentlich* bezeichnet man Beobachtungsverfahren, bei denen gegenüber den Beobachteten die Tatsache, daß sie beobachtet werden, so weit wie möglich verborgen oder kaschiert wird.

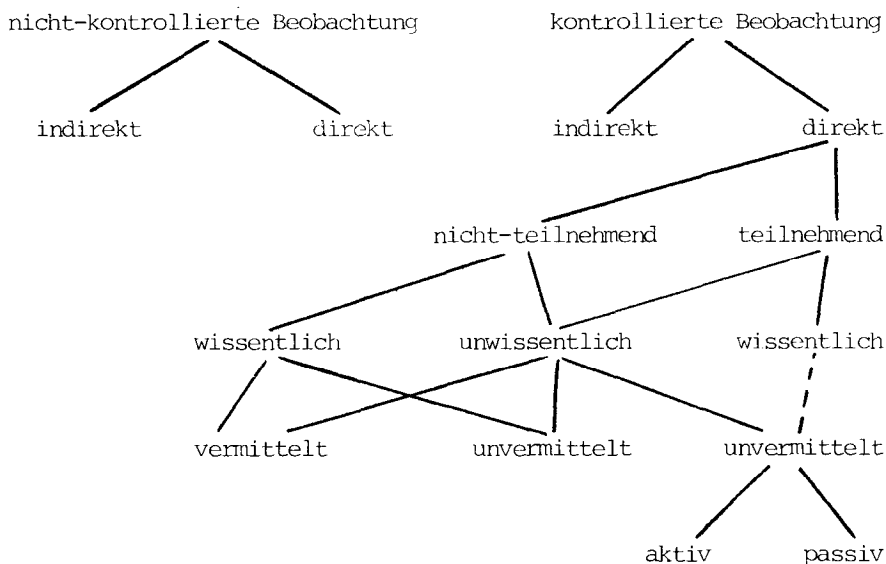


Abb. 1: Schema der Methoden der Verhaltensbeobachtung (nach Graumann, 1966)

Die herkömmliche Unterscheidung zwischen *freier* und *systematischer* Beobachtung greifen v. Cranach & Frenz (1969, S. 269) auf. Während freie Beobachtungen „ohne methodische Einschränkung“ vorgenommen werde, geschehe bei der systematischen Beobachtung die „Anwendung von Beobachtungssystemen als Meßverfahren, welche die quantitative Erfassung operational definierter Variablen mit angebbarer Objektivität, Zuverlässigkeit und Gültigkeit erlauben“.

2.2 Teilnehmende Beobachtung

Da es für Verhaltensbeobachtung wichtig ist anzugeben, auf wen das Verhalten des Beobachteten gerichtet ist, muß die Rolle des Beobachters in der Beobachtungssituation untersucht werden, insofern und insbesondere als er auch für die beobachteten Personen Teil ihrer Situation sein kann. Diese Problematik ist bei der Entwicklung der *Methode der teilnehmenden Beobachtung* (participant observation) durchweg gesehen worden (Whyte 1953 als Klassiker, Schwartz & Schwartz 1955, Kluckhohn 1956, McCall & Simmons 1969, Friedrichs & Lüdtke 1973). Fälle, in denen schon die bloße Anwesenheit eines Beobachters zu anderem Verhalten führte als in Situationen ohne einen Beobachter, sind schon früh berichtet worden, etwa von Polansky et al. (1949) bei verhaltensauffälligen Kindern. Die Frage der Reaktivität ist in jüngster Zeit ausführlicher bei der Erforschung von Verhaltenseinschätzungen untersucht worden und soll im Kapitel von Feger und Graumann dargestellt werden.

Teilnehmende Beobachtung verlangt in der Regel jedoch mehr als bloße Anwesenheit, oft volle Mitgliedschaft des Beobachters in der untersuchten sozialen Gruppierung. Dies wird verständlich, wenn man sich die hauptsächlichsten Anwendungsbereiche dieser Methodik vergegenwärtigt: primitive Kulturen, Subkulturen, Kommunen, Krankenhäuser, Gefängnisse, Fabriken, Bürokratien, Militär, Kulte, Familien, Verbrecherbanden. Grümer (1974) verweist in Anlehnung an Kunz auf drei nach der Rolle des Sprachsystems unterschiedene Forschungssituationen, in denen die teilnehmende Beobachtung eingesetzt werden könne: Erstens Situationen, in denen Beobachter und Beobachtete über ein unterschiedliches Sprachsystem verfügten, wie etwa in der Verhaltensforschung an Tieren; zweitens Situationen, in denen Beobachter und Beobachtete zwar ein unterschiedliches Sprachsystem haben, jedoch wenigstens teilweise lernen können, eine gemeinsame Sprache zu benutzen, beispielsweise bei ethnologischen Untersuchungen in fremdsprachigen Kulturen; drittens Forschungssituationen, in denen zwar ein gemeinsames Sprachsystem besteht, jedoch Unterschiede und Abweichungen vorkommen, die das Verständnis eines Beobachters für eine Situation beeinträchtigen können, z.B. bei Untersuchungen mit Kleinkindern, Geisteskranken und kulturellen Subgruppen.

McCall & Simon (1969, S. 3) beschreiben, was unter teilnehmender Beobachtung bei ihrer Anwendung faktisch verstanden wird:

„.... *is* most sensibly regarded, operationally, as the blend of methods and techniques that is characteristically employed in studies of social situations or complex social organizations of all sorts. These are studies that involve repeated, genuine social interaction on the scene with the subjects themselves as a part of the data-gathering process.“

An dieser Umschreibung sind uns zwei Punkte wichtig, auf die wir weiter eingehen wollen, zum einen der Hinweis, daß teilnehmende Beobachtung typischerweise eine Kombination von Methoden darstellt, und daß sie den Beobachter in eine soziale Interaktion einbindet. Das Wesen der Methode (und die Quelle ihrer Schwierigkeiten) besteht darin, daß der Beobachter im sozialen Feld eine bestimmte Rolle zu spielen hat, und somit nicht nur eine ohnehin oft schwierig als solche zu definierende „repräsentative“ Auswahl des Feldes den Wert der Beobachtungen bestimmt, sondern auch das Geschick, mit dem der Beobachter eine ihm angemessen erscheinende Rolle wählt und diese spielt. Die Rolle muß vom Beobachter so übernommen werden, oder es muß von denen, die solche Rollen gewöhnlich spielen, so ein geeigneter Beobachter rekrutiert werden, daß das soziale Feld nicht verfälscht wird. Ob dies gelungen ist, kann man bisweilen aus den Korrelationen mit Daten erschließen, die mit Hilfe anderer Verfahren, insbesondere durch unwissentliche Beobachtung gewonnen wurde, oder - für den typischen Anwendungsbereich realistischer - aus Übereinstimmung von eigenen Beobachtungen mit den Berichten von Informanten. (Zu spezifischen Problemen z.B. der Informantenauswahl, der Generalisierbarkeit der Befunde, möglicher Veränderung des Kategoriensystems des Beobachters bei langdauernder Interaktion und der Prüfung der internen Validität siehe insb. McCall & Simons und die dort abgedruckten Artikel).

3. Die Planung von Beobachtungen

Bei der Planung muß zunächst entschieden werden, was zu beobachten ist. Für diese Entscheidung muß das Universum der Beobachtungen definiert, die Einheit der Analyse bestimmt und das Kategoriensystem entwickelt werden. Danach sucht man Antworten auf die Frage, welche der unter diesen Vorgaben meist zahlreichen möglichen Beobachtungen tatsächlich realisiert, wie also eine Auswahl aus den Universen von Personen, Situationen, Zeitpunkten etc. getroffen werden soll.

3.1 Das Universum von Beobachtungen

Um die für die folgenden Darlegungen maßgebliche *Facettentheorie* (Guttman 1959, 1971; im folgenden nach Borg, 1977) einzuführen, skizzieren wir zunächst eine Studie von Guttman & Guttman (1976), die sich von den meisten Anwendungen der Facettentheorie dadurch unterscheidet, daß sie sich nicht auf die Konstruktion von Test- oder Fragebogenitems und der Erklärung ihrer korrelativen Zusammenhänge bezieht, und daß sie eine Sekundäranalyse bereits vorliegender Beobachtungen in mehreren Experimenten zu Streßindikatoren bei Mäusen darstellt. Facettentheoretische Analysen von Beobachtungen

finden wir auch bei Canter (1977, 1977b) im Bereich der Architektur- und Umweltpsychologie.

Bei der Durchsicht einer Reihe von Untersuchungen zu emotionalem Verhalten von Mäusen stellten Guttman & Guttman fest, daß die Tiere unter verschiedenartigen Bedingungen beobachtet worden waren. Diese verschiedenen experimentellen Situationen wie „offener Raum“, „Lauftrad“, Schwimmlabyrinth“ usw. stellen verschiedene Elemente einer *Facette* „experimentelle Situationen“ dar, wobei von der Facette vermutet wird, daß ihre Elemente zu Streß in unterschiedlichem Ausmaß führen. Die Beobachtungshinsicht in der untersuchten Forschung waren Maße der Frequenz, der Latenz und der Dauer. Da sich nach Ansicht von Guttman & Guttman das Ausmaß des Streß in diesen „zeitlichen Merkmalen“ unterschiedlich zeigen kann, stellen diese Merkmale eine weitere Facette dar. Schließlich zeige sich Streß in Verhaltensweisen, die entweder der willkürlichen Kontrolle unterliegen, wie z. B. Dauer des Verbleibens im offenen Raum, oder vom autonomen System gesteuert werden, z.B. Harn lassen und Koten. Die Art der Kontrolle stellt somit eine dritte Facette dar. Variation innerhalb einer Facette könne zu hohen oder niedrigen Ausprägungen des Streß führen; diese abhängige Variable stellt die durch die Facetten zu erklärenden Beobachtungen dar, Beobachtungen, an denen hier das unterschiedliche Ausmaß an Streß interessiert.

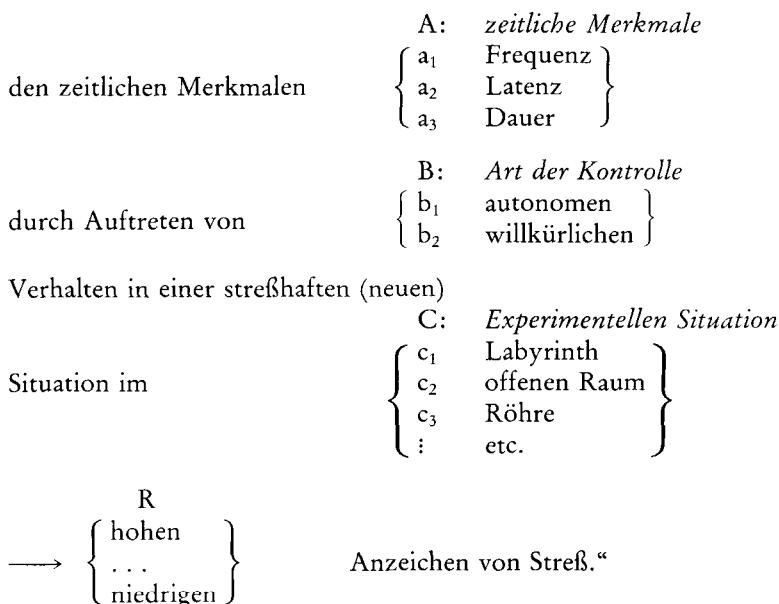
Verallgemeinert man dieses Beispiel zu dem „general paradigm of empirical research in the social sciences“ (Borg, 1979, S. 65), dann sieht man, daß drei Komponenten analytisch unterschieden werden: Eine Population P, die mit einer Menge Reize S konfrontiert wird und auf diese mit den Reaktionen R antwortet. P stellt in der Methodik der wissenschaftlichen Beobachtung die zu beobachtenden Menschen und Tiere dar, S die Variablen der Beobachtungssituation, R die für die theoretische oder praktische Fragestellung interessierenden Verhaltensweisen. Die Beziehung dieser drei Komponenten zueinander läßt sich als *Abbildungssatz* (mapping sentence) schreiben:

$$P \times S \rightarrow R$$

Das kartesische Produkt $P \times S$ stellt den Definitionsbereich (domain), R stellt den Bildbereich (range) dar; der Pfeil symbolisiert die Abbildung von $P \times S$ in R und kann als „ist verknüpft mit“ gelesen werden. Die Festlegung von Definitions- und Bildbereich stellt das *Universum von Beobachtungen* (Universe of observations) her. Eine konkrete Beobachtung besteht dann - wenn man sie von ihrer Leistung her charakterisiert - in der Zuordnung von Elementen des Definitionsbereiches zum Bildbereich.

Im einfachsten Fall bestehen P, S und R jeweils aus nur einer Menge (Facette), jedoch kann jede Komponente selbst ein kartesisches Produkt sein. Im Beispiel lautet der Abbildungssatz:

„Das Ausmaß, in dem Tier (X) Streß zeigt in



In diesem Beispiel hat S drei Facetten: A, B und C. P, die Variable „Tier (X)“, ist nicht weiter ausgeführt. Dies hätte - wären die Forscher daran interessiert gewesen und hätten entsprechende Arbeiten für die Sekundäranalyse vorgelegen - leicht geschehen können, beispielsweise indem eine Facette verschiedene Zuchtstämme von Ratten unterscheidet, die auf Streß unterschiedlich reagieren.

Ein wesentlicher Gewinn bei der Anwendung der Facettentheorie zur Planung und Analyse wissenschaftlicher Beobachtungen kann in der hierbei geforderten Definition der Facetten bestehen, weil so die tatsächlich berücksichtigte oder mögliche Gesamtheit von Beobachtungen angegeben wird. Daraus können nicht nur Vorschriften für das Ziehen der Stichproben (z.B. für jede Facettenkombination mindestens eine Beobachtung) abgeleitet und begründet werden, vielmehr wird oft erst auf diese Weise eine genaue Formulierung der Fragestellung erzwungen und der Bereich der Verallgemeinerbarkeit festgelegt. Letzteres behandelt die später besprochene Generalisierbarkeitstheorie. Die Prüfung der konvergierenden und diskriminierenden Validität, die später ebenfalls dargestellt wird, stellt facettentheoretisch die Analyse der Facetten „Inhalte“ und „Methoden“ der untersuchten Variablen dar. Ferner kann man bestimmte Prinzipien heranziehen, um die Struktur im Bildbereich aus der Struktur des Definitionsbereiches vorherzusagen. In der erwähnten Reanalyse konnten Guttman & Guttman beispielsweise zeigen, daß die Höhe der Korre-

lationen zwischen den verschiedenen Streßindikatoren davon abhängt, hinsichtlich wievieler und welcher Facetten die Beobachtungen Gemeinsamkeiten aufwiesen. Diese Struktur der korrelativen Beziehungen zwischen den verschiedenen Indizes, dargestellt mit Hilfe einer multidimensionalen Skalierung, erwies sich außerdem als über die verschiedenen Experimente hinweg vergleichbar.

Wir besprechen nun einige in Psychologie und Nachbarwissenschaften häufige spezielle Fälle der Konkretisierung der Mengen P, S und R. Bei der facetten-theoretischen Betrachtung von Tests, Einstellungs- und Persönlichkeitsfragenbogen werden die Verhaltensrealisationen an verschiedenen Personen betrachtet, die nach demographischen oder zuvor erhobenen differentialpsychologischen Merkmalen geordnet sein können oder lediglich durch ihre im Namen festgehaltene Identität als unterscheidbare Elemente dieser einen oder aller personenbeschreibenden Facetten festgestellt werden. S besteht aus Items, die beispielsweise in einem Intelligenztest so in mehrere Facetten geordnet werden könnten, daß der Guilfordsche Würfel Pate steht. Beobachtet wird, wie eine Person auf ein Item reagiert, dabei können im Bildbereich z.B. bei Leistungstests sowohl Richtigkeit als auch Schnelligkeit als auch Qualitätsmerkmale der Antwort spezifiziert werden.

Eine Facette kann auch nur ein Element aufweisen. Dies wird in der Unterscheidung verschiedener Strategien zum Ziehen von Verhaltensstichproben deutlich (s.u.). Wenn der Merkmalsträger zugleich der Beobachter ist, dann haben wir den Fall der *Selbstbeobachtung* vor uns. Der Beobachter beschreibt dabei, sei es gegenüber sich selbst, oft gegenüber anderen, dem Untersucher etwa, sein Erleben und Verhalten, z.B. seine Reaktionen auf verschiedene Farben. Im Abbildungssatz können dabei durchaus mehrere Personen als Elemente einer Facette vorgesehen sein, etwa vollständig (completely crossed) oder nur teilweise (nested) verschiedenen Untersuchern zugeordnet. Wesentlich ist, daß bei dieser Beobachtungsart die Basis für über Personen verallgemeinernde Aussagen in Äquivalenzannahmen bestehen muß, etwa der Art: Zwei Personen meinen das Gleiche, wenn sie beim Vergleich zweier beleuchteter Flächen eine als „heller“ bezeichnen. Die Menge der Merkmalsträger kann selbst als kartesisches Produkt $P \times P$ geschrieben werden, wenn beispielsweise die (*sozialen*) *Beziehungen* zwischen Personen, Tieren, Gruppen etc. beobachtet werden sollen. Bei Plänen mit wiederholten Beobachtungen kann man in den Definitionsbereich die Facette „*Zeitpunkt*“ aufnehmen; Levy & Guttman (1975) berichten eine facettentheoretisch angelegte Untersuchung mit fünf wiederholten Messungen.

Beobachtungspläne unterscheiden sich unter anderem danach, wie expliziert und vorstrukturiert Definitions- und Bildbereich durch vorausgegangene wissenschaftliche Arbeit sind. Im einen Extremfall, in der Regel bei Datengewinnung in einem *Experiment*, ist vor Beginn der Beobachtung festgelegt, wie alle

Facetten definiert sind, d.h. welche Elemente sie enthalten oder enthalten können. Im anderen Extrem, oft in der *Feldbeobachtung*, besteht bisweilen primär die wissenschaftliche Leistung darin, überhaupt „Beobachtenswertes“ zu finden, d.h. hier: die Facetten zu definieren und einander zuzuordnen. Noch einige Querverweise, um zu erläutern, in welcher Hinsicht wir das facettentheoretische Vorgehen für fundamental halten: Einige Arbeitsrichtungen in der Psychologie befassen sich damit zu bestimmen, mit welchen und wie vorstrukturierten Mengen von Kategorien alltägliche und wissenschaftliche Beobachtungen angestellt werden. Forschungen zur „impliziten Persönlichkeitstheorie“ ermitteln u.a., welche Elemente der Definitionsbereich bei der Beschreibung von Personen enthält, und wie die multivariate Struktur dieses Bereiches sich aus der Kombination der Facetten ergibt. Nosologische Klassifikationssysteme der psychologischen und psychiatrischen Diagnostik sollten m.E. als Abbildungssätze geschrieben werden, so daß Symptomkorrelationen erklärbar erscheinen. Einstellungstheorien (z.B. Fishbein & Ajzen 1975, Feger 1979) spezifizieren verschiedene Mengen von Komponenten und deren Bewertung, wobei z.T. die Vpn selbst die Elemente der Menge generieren.

3.2 Bestimmen der Beobachtungseinheit

Die Genauigkeit, mit der die Komponenten im Abbildungssatz festgelegt werden können, ist nicht unbegrenzt. Oft tritt an die Stelle einer Menge in S ein ausgearbeitetes Kategoriensystem. In ihm, in der Instruktion und in der Beobachtungsschulung versucht der Forscher, die Beobachtungseinheit so präzise wie möglich festzulegen. Dabei tritt zu der inhaltlichen Bestimmung dessen, was beobachtet werden soll - die meistens im Kategoriensystem vorgegeben wird - der Versuch des Forschers, vor allem die zeitliche Dauer, und auch den räumlichen Bereich einer einzelnen Beobachtung festzulegen. Der Forscher versucht, den Beobachter dazu zu bringen, den Verhaltensstrom in möglichst eindeutig angebbarer Weise in Zeitabschnitte zu gliedern. Schließlich will ein Forscher oft nicht nur wissen, was beobachtet wurde, sondern auch, wann und wo. So sehr sich ein Forscher auch bemüht, den Beobachter in dieser Hinsicht so genau wie möglich anzuleiten, er stößt doch auf psychologische Grenzen des Beobachters, die im menschlichen Wahrnehmungssystem begründet sind. Hier überschneiden sich eine allgemeine Methodenlehre der Beobachtung und die Wahrnehmungspsychologie; einige Aspekte dieser Thematik behandeln Feger & Graumann (in diesem Band).

Wie nun im Zusammenspiel von Forscher und Beobachter - auch wenn eine Person beide Rollen spielt - die Definition der Beobachtungseinheit ausgehandelt wird, hat für die Beobachtungsergebnisse weitreichende Folgen. Aus

Arbeiten des Kreises um Newton (1973, 1976, Newton & Enquist 1976, Newton et al. 1977; s.a. Feger & Graumann in diesem Band) wissen wir, daß Beobachter spontan selbst das Beobachtete in Einheiten untergliedern. Daß sich unterschiedliche Ergebnisse einstellen können, wenn entweder die zeitliche Ausdehnung vorgegeben oder die Dauer des beobachteten Phänomens vom Beobachter selbst festgelegt wird, berichten Hayes et al. (1970). Einige Autoren betonen nun, man solle die Wahl der Einheit nicht dem Beobachter überlassen, so z.B. Grümer (1974, S. 41): „Die Definition von Beobachtungseinheiten wird in strukturierten Beobachtungen nicht dem Beobachter selbst überlassen bleiben können, sondern wird zu einer spezifischen Aufgabe eines Forschers oder Untersuchungsleiters werden müssen. Aufgrund seiner Kenntnisse über Strukturen und Verhaltensabläufe in einem Beobachtungsfeld ist er in der Lage, Beobachtungseinheiten festzulegen.“ Und so legen auch v. Cranach & Frenz (1969, S. 286) in ihrer Definition fest: „Als Beobachtungseinheit wird derjenige Bestandteil in einem Verhaltensablauf bezeichnet, der dem Untersucher als kleinstes, nicht reduzierbares Ereignis zur Analyse des Verhaltens notwendig erscheint.“ Die Notwendigkeit, die Einheit durch den Forscher vorzugeben, wird mit den weitreichenden Folgen begründet, die sich aus der Wahl der Einheit ergeben: „Einmal legt man mit der Enge bzw. Breite der Beobachtungseinheit zugleich die Variationsmöglichkeiten fest, die das Beobachtungssystem bei der Abbildung des realen Geschehens zuläßt. Zum anderen muß man alle Fälle, die einer Beobachtungskategorie zugeordnet worden sind, als untereinander invariant ansehen.“ (v. Cranach & Frenz, 1969, S. 286).

Ferner ist die Wahl der Einheit für die später behandelte Frage der Beobachterübereinstimmung relevant. Idealerweise stimmen Beobachter dann überein, wenn sie Gleiches beschreiben. Die Beobachtungseinheit festzulegen bedeutet unter diesem Gesichtspunkt einen Versuch, zur Beobachtung des Gleichen anzuregen. Möglicherweise gibt die zu untersuchende Fragestellung vor, wie aus theoretischen oder praktischen Gründen die Einheit gewählt werden muß. Durch die Wahl der Einheit wird das „Auflösevermögen eines Beobachtungsansatzes“ (Faßnacht, 1979) festgelegt. Das, was bei der Beobachtung als Einheit zusammengefaßt und beschrieben wird, kann in der Datenanalyse nicht mehr differenziert werden. Bales (1950) gibt zur Beschreibung von Interaktion in Problemlösegruppen zwölf globale Kategorien vor, wodurch umfassende Einheiten gebildet werden dürften, während Frey & Pool (1967) allein 125 verschiedene Kopfstellungen einer sitzenden Person unterscheiden. Man sollte jedoch festhalten, daß die Beobachtungseinheit letztlich vom Beobachter gebildet wird, wenn auch sein Spielraum dabei durch genaue Instruktion, Training u.a. eingeschränkt werden kann. Die Übereinstimmung bei der Segmentierung sollte gegebenenfalls geprüft werden. Die Grenze des Auflösevermögens der Beobachter ist erreicht, wenn die Verlässlichkeit der Unterscheidungen zu gering wird.

Wie Einheiten gebildet werden, unterscheiden Kalbermatten & v. Cranach (1980) nach vier Gesichtspunkten. Sie akzentuieren „*natürliche*“ gegenüber „*künstlichen*“ Einheiten. Natürliche Einheiten sind „bereits vor dem wissenschaftlichen Prozeß vorhandene Unterscheidungstendenzen des Beobachters“, z.B. ob eine Person lacht oder nicht. Künstliche sind „theoretisch fundierte, im Forschungsprozeß gewonnene Merkmale“, ob eine Person beispielsweise die Wangenmuskeln aktiviert. Zweitens wird zwischen *sozial bedeutungsvollen* und *physikalisch definierten* Einheiten unterschieden. Bei ersteren werden die Bedeutungen von der sozialen Gemeinschaft zuerkannt. Sie lassen sich feststellen an der Übereinstimmung, mit der Mitglieder der Gemeinschaft sie definieren, und ihre Funktion sei es, den „mehr oder minder reibungslosen Ablauf der Interaktion zu gewährleisten“. Drittens werden *funktionale* gegenüber *strukturellen* Einheiten unterschieden. Funktionale seien „dynamisch konzipiert“, sie hielten „die wechselseitige Auswirkung einer Einheit auf eine andere und auf das Gesamtergebnis des Verhaltens fest“. Drohen wird als Beispiel genannt; die geballte Faust wäre ein Beispiel für strukturelle Einheitsbildung. Viertens wird, wie auch sonst häufig in der Literatur, zwischen *mola-*ren und *molekularen* Einheiten unterschieden, nach der „Höhe des Abstraktionsgrades“ oder dem Umfang der Zusammenfassung.

Kalbermatten & v. Cranach vertreten dann die Auffassung, bei der Analyse menschlicher Handlungssysteme sei es erforderlich, verschiedene, hierarchisch geordnete Organisationsebenen der Einheiten zu unterscheiden. Jede Einheit einer höheren Ebene lasse sich in Untereinheiten der nächsttieferen Ebene aufgliedern; die Einheiten verschiedener Ebenen unterschieden sich auch qualitativ, da die höheren Einheiten in der Regel von ihrer Funktion her und durch soziale Bedeutungsverleihung definiert seien. Eine wesentliche Aufgabe der Forschung bestehe dann in der Untersuchung des genauen Zusammenhanges zwischen Einheiten verschiedener Ebenen (z.B.: Welche Einheiten der tieferen Ebene - etwa Ballen der Faust und Heben der Stimme - werden wie kombiniert, damit die Einheit „Drohen“ entsteht?).

3.3 Kategoriensysteme

Der Bildbereich eines Abbildungssatzes legt fest, welcher Aspekt am Beobachteten untersucht werden soll. Im einfachsten Fall, wie im Streßbeispiel, ist dies lediglich *eine* Variationsdimension, eine Hinsicht, nämlich das Ausmaß des Stresses, das sich in verschiedenen Verhaltensweisen unter verschiedenen Bedingungen zeigt. Eine Kategorie beschreibt Ausprägungen von Reaktionsmerkmalen, die qualitativ oder quantitativ sein können. Variation kann unter den gleichen Bedingungen und an den gleichen Verhaltensweisen in mehr als einer Hinsicht unterschieden und zugleich beobachtet werden. Die Auflistung dieser Hinsichten in ihren qualitativen und quantitativen Abstufungen stellt

das Kategoriensystem dar. Ein Kategoriensystem stellt also die Unterscheidungen zusammen, die der Beobachter treffen kann, und legt fest, bei welchen Zuständen und Prozessen am Beobachteten er sie treffen soll. Wir sprechen von System, weil die Unterscheidungsmöglichkeiten mehr oder weniger geregelt aufeinander bezogen sind, etwa durch die Vorschrift, das Beobachtete solle nur einer Abstufung von mehreren bei einer Kategorie zugeordnet werden können. Zum Kategoriensystem i.e.S. kommen häufig noch Anweisungen hinzu, kodifizierte Anleitungen an den Beobachter, wie er das Beobachtete erfassen, zuordnen und registrieren soll, was das Kategoriensystem i. w. S. ausmacht.

V. Cranach & Frenz (1969) wie auch Medley & Mitzel (1963) unterscheiden drei Arten von „Beobachtungssystemen“, und zwar Zeichen-Systeme, Kategorien-Systeme und Schätzskalen (rating scale format). Ein einfaches Zeichensystem bestünde beispielsweise darin, daß ein Beobachter immer dann, wenn eine Person bei Rot eine Ampel passiert, einen Registrierknopf drückt. Das Überschreiten der Markierungslinie wäre für ihn ein Zeichen, alles nicht als Zeichen definierte Verhalten bleibt unregistriert. Eines der bekanntesten Kategoriensysteme im Sinne von Medley & Mitzel ist das von Bales (1950, 1968; siehe dazu auch Grümer, 1974, Manz, 1974). In der Regel hat der Beobachter bei diesen Kategoriensystemen nicht nur - wie bei Zeichensystemen - zu entscheiden, ob ein bestimmtes Verhalten aufgetreten ist, sondern auch welches von den einander ausschließenden Möglichkeiten des meist als vollständig konzipierten Kategoriensystems.

Weitere Beispiele für Kategoriensysteme sind das von Caldwell (1969), das sich auf Verhalten von Kindern bezieht, oder das von Kaufman & Rosenblum (1966) für das Sozialverhalten von Primaten; Simon & Boyer (1974) geben eine Übersicht über 99 Systeme. Ausgearbeitete Systeme von Ratingskalen finden sich z.B. in Thomae (1968, s. auch Rudinger & Feger, 1970).

Viele Systeme sind keine reinen Zeichen-, Kategorien- oder Schätzskalensysteme. Für ihre Herleitung sind selten theoretisch so konsistent systematisierte Annahmen wie bei Bales vorhanden, doch sollte der Weg von der Fragestellung des Forschers zur Auswahl und Formulierung der Kategorien so einsichtig und zwingend wie möglich sein; die Facettentheorie könnte für die Ableitung von Kategoriensystemen verwendet werden. Viele Kategoriensysteme versuchen in dem Sinne vollständig zu sein, daß alles beobachtete Verhalten einer Kategorie zugeordnet werden kann, und sei es einer Restkategorie, in die alles nicht klassifizierbare eingewiesen wird. Dadurch entstehen logische Abhängigkeiten zwischen den Kategorien, die die statistische Auswertung erschweren. Wenn von zwei Kategorien Kategorie A als „soziales Verhalten“ und B als „nicht-soziales Verhalten“ definiert sind, muß sich perfekter negativer Zusammenhang ergeben. Diese, in umfangreichen Kategoriensystemen

nicht so leicht erkennbaren logischen Zusammenhänge vermischen sich mit den empirischen und müssen bei der Auswertung berücksichtigt werden. Umfangreiche Kategoriensysteme führen bei unabhängiger Auswertung jeder Kategorie oft zu einem erhöhten faktischen α -Niveau der statistischen Signifikanztests, weshalb multivariate Verfahren häufiger als bisher verwendet werden sollten. Ein weiteres Problem ergibt sich bei Systemen, die wie eine check-list angelegt sind, d.h. vom Beobachter lediglich zu registrieren verlangen, wie häufig eine Verhaltensweise auftritt. In der Interpretation wird dann bisweilen unterstellt, die Verhaltensdauer stünde mit der Häufigkeit in direkter proportionaler Beziehung. Für diese Unterstellung fand zwar Adams (1970) bei Schülerverhalten im Unterricht positive Evidenz, aber eine allgemeine Austauschbarkeit von Dauer- und Häufigkeitsmaßen gibt es nicht.

Longabaugh (1980) unterscheidet, was er „Kodiersysteme“ nennt, (1) danach, ob sie aus einer Theorie abgeleitet sind oder auf nicht systematisierten Erfahrungen mit dem Gegenstandsbereich aufbauen, (2) nach der Breite gegenüber der Detailliertheit, mit der sie einen Bereich erfassen, und (3) nach dem Ausmaß, in dem Schlußfolgerungen, Interpretationen durch den Beobachter erforderlich sind: „... to what extent is the coder required to take advantage of the fact that he is a socialized human being, sharing a common culture, in order to characterize the behavior?“ Auf die Folgerungen, die sich aus diesem dritten Gesichtspunkt für die Reliabilitätsfrage ergeben, gehen wir später ein.

Um einzelne Kategorien und Kategoriensysteme zu klassifizieren, scheint es z.B. für die Diskussion ihrer Validität wie für das Training der Beobachter günstig, danach zu fragen, welche Aufgaben sie dem Beobachter stellen. Da Beobachtung ein kontrollierter Prozeß der Wahrnehmung ist, sind die Aufgaben grundsätzlich diejenigen von Wahrnehmung allgemein, die man mit Luce und Galanter (Luce, 1963, Luce & Galanter, 1963a, b) bezeichnen kann als Entdecken (detection), Wiedererkennen (recognition), Unterscheiden (discrimination) und Quantifizierung (scaling). Diese vier Aufgaben lassen sich, wie Luce (1963, S. 105) dies für eine psychophysikalische Behandlung von Wahrnehmungsproblemen getan hat, als vier Fragen formulieren: (1) Liegt das zu beobachtende Verhalten vor? (2) Welche von mehreren möglichen Verhaltensweisen liegt vor? (3) Unterscheidet sich dieses Verhalten von jenem? und (4) Wie verschieden ist dieses Verhalten von jenem? Selbst bei den einfachsten Kategoriensystemen sind Aufgaben des Entdeckens und des Wiedererkennens zu lösen, bei Kategoriensystemen im Sinne von Medley & Mitzel auch Diskriminationsaufgaben, und bei Schätzskalen zusätzlich Quantifizierungsaufgaben. Um diese Aufgaben lösen zu können, definiert der Forscher für den Beobachter durch Instruktion, Kategorien und Schulung **Zuordnungsregeln**, die möglichst explizit vorschreiben, welche Erscheinungen am Beobachtungsgegenstand welchen Kategorien zuzuordnen sind. Duncan & Fiske (1977, S. 15) sprechen in diesem Zusammenhang von recognition rules: „The rules for

the identification of events as instances of the applicable categories will be termed *recognition rules*. Using the investigator's category system, the task of the rater is in the first place to recognize the occurrence of an event specified by a recognition rule.“ Reliabilität und Validität von Beobachtungen hängen davon ab, wie explizit die Zuordnungsregeln formuliert, angewendet und wie gut ihre Anwendung geprüft werden kann.

3.4 Auswahlen aus dem Universum der Beobachtungen

Meistens können nicht alle Beobachtungen, die gemäß dem Abbildungssatz prinzipiell möglich wären, auch tatsächlich angestellt werden, z.B. weil ihre Zahl zu groß ist, oder Kosten und Zeitaufwand zu hoch wären. Deshalb muß fast immer eine Auswahl getroffen, und diese begründet werden. Um Auswahlarten zu beschreiben, ist es hilfreich, den facettheoretischen Begriff des *Struktupels* einzuführen. Da die Elemente aller Facetten sich vollständig kreuzklassifizieren lassen, d.h. aus jeder Facette jedes Element mit jedem Element jeder anderen Facette kombiniert werden kann, entstehen Kombinationen mit k Elementen, wobei k die Zahl der Facetten ist. Ihre Anzahl entspricht dem kartesischen Produkt der Facetten. Jede Kombination stellt ein Struktupel dar, in dem das erste Element aus der ersten, das zweite aus der zweiten usw. Facette stammt. Die erste Möglichkeit auszuwählen besteht darin, nicht für jedes Struktupel eine Beobachtung zu realisieren, also aus den Struktupeln auszuwählen. So hat beispielsweise Jordan (1971) bei der Konstruktion eines Vorurteils-Fragebogens nicht alle Kombinationen ausgewählt, weil sie seines Erachtens teilweise keinen psychologischen Sinn ergaben. In einem solchen Fall könnte man die Facetten nicht als vollständig überkreuzt, sondern als geschachtelt ansehen. Kann man das nicht, führt die Auswahl aus Facetten zu einer Einschränkung der Allgemeingültigkeit der Aussagen. Im Extremfall untersucht man nur ein Element einer Facette, z.B. berufstätige Frauen aus der Personen-Facette. Eine zweite Möglichkeit auszuwählen stellt die Auswahl *innerhalb* eines Struktupels dar. Bei Testitems etwa gibt es eine Fülle von Formulierungen, in denen der gleiche Inhalt ausgedrückt werden kann, und eine Auswahl ist aus Aufwandgründen erforderlich. Alle Realisationsmöglichkeiten, die durch das gleiche Struktupel beschrieben werden, sind für die Theorie, die zum Abbildungssatz geführt hat, äquivalent. Je mehr Realisationen dann pro Struktupel vorliegen, um so verlässlicher ist die Datenbasis.

3.4.1 Auswahl von Personen

Es gibt verschiedene Gründe, warum man mehr als eine Vp untersucht, selbst wenn man überzeugt ist, das interessierende Verhalten könne in allen Erschei-

nungsformen und aufgrund aller relevanten Bedingungen an jeder beliebigen und somit jeder einzelnen Person beobachtet werden. (1) Wenn der Forscher das interessierende Verhalten nicht herbeiführen kann oder will, oder es zu selten bei einer Person auftritt, beobachtet er mehrere, um die *Auftretenswahrscheinlichkeit* des Verhaltens zu erhöhen. (2) Der Forscher möchte eine Vp unter verschiedenen Bedingungen beobachten, jedoch kann die gleiche Vp nur unter einer Bedingung untersucht werden, z.B. wegen störender carry-over-Effekte (*Kontrollproblem*). Dann ergibt sich insbesondere in Feldstudien das Problem, unter verschiedenen Bedingungen solche Personen zu beobachten, die vergleichbar sind, so daß Verhaltensunterschiede eindeutig den Bedingungen zugeordnet werden können (interne Validität). (3) Der Forscher möchte Aussagen treffen über die *Generalität*, über Verbreitungsgrad und interindividuelle Variabilität des Verhaltens. Nur bei dieser Absicht werden Überlegungen relevant, wie man eine repräsentative Stichprobe aus der interessierenden Population gewinnen kann (dazu gute Einführungen: Kish, 1953; Scott & Wertheimer, 1962; zur Vertiefung: Böltken, 1976; Cochran, 1953; Deming, 1950; Hansen et al., 1953; Yates, 1953).

Ein besonderes Problem ergibt sich aus der Tatsache, daß ‚freiwillige Vpn‘ untersucht werden, was oft schon aus ethischen Gründen unvermeidlich ist. Die Arbeiten von Rosenthal (1965) sowie Rosenthal & Rosnow (1969, 1975; dort weitere Literatur) legen zwei Schlüsse nahe: Es gibt bestimmte situative Umstände, die einige Personen zur freiwilligen Vp werden lassen, andere Personen nicht. Und zwischen freiwilligen und zwangsrekrutierten Vpn gibt es einige Unterschiede in Merkmalen wie Höhe der Schulbildung, des sozialen Status, der Intelligenz, der Anpassung und motivationaler Bedürfnisse. Das Problem ist damit im Sinne von Campbell & Stanley (1963) eines der externen Validität, der fraglichen Grenzen der Verallgemeinerbarkeit von Befunden. Allerdings weist Kruglanski (1975) darauf hin, daß Versuchsergebnisse nicht generell von der Freiwilligkeit der Vpn abhängen. Bredenkamp (persönliche Mitteilung) weist darauf hin, nur wenn eine solche Abhängigkeit durchgängig gegeben sei, müsse man diese Artefaktmöglichkeit auch allgemein kontrollieren. Hingegen müßte zum speziellen Nachweis eines Artefaktes gezeigt werden, daß der Faktor „Freiwilligkeit“ disordinal mit der unabhängigen Variablen einer Untersuchung derart interagiert, daß die Relation zwischen unabhängiger und abhängiger Variable für unterschiedliche Gruppen verschieden ausfällt. Die bei Rosenthal & Rosnow aufgeführten Untersuchungen, die diese Interaktion geprüft haben, zeigen eben nicht disordinale Interaktionen (s. Bredenkamp 1980).

Wenn Verhaltensweisen untersucht werden sollen, die von anderen Personen nicht beeinflußt werden, oder bei denen ein solcher Einfluß nicht interessiert, kann man die Stichprobe so ziehen, daß jede Person unabhängig von jeder anderen eine gleiche oder bekannte Chance hat, gezogen zu werden. Sonst

kann das Paar, die Gruppe, die Kultur als Element der Population gezogen werden, wobei die Facettentheorie zur Populationsdefinition benutzt werden kann. Für kulturvergleichende Studien ringen zwei Kriterien für die Zusammensetzung der Personenstichprobe um gleichzeitige Beachtung „. . . whether a given relationship between variables will be obtained across cultures (irrespective of people) and across people (irrespective of their cultural membership)“, Longabaugh (1980, S. 73). Eine bemerkenswerte Lösung dieser Aufgabe findet sich in Whiting et al. (1966).

3.4.2 Auswahl und Schulung von Beobachtern

Selten hat ein Untersucher überhaupt die Möglichkeit, aus einer Gruppe von mehreren die ihm geeignet erscheinenden Beobachter auszusuchen. Die formalen Kriterien sind bei Auswahlentscheidungen meistens hinreichend hohe Übereinstimmung mit einem Standard, mit dem „mittleren“ oder typischen Beobachter (s.u. Krippendorff, 1970), oder eine hohe oder nicht mehr zu steigernde Übereinstimmung zwischen einem Paar oder einem Team aus der Gesamtmenge der Beobachter. Bock (1956) hat die Auswahl von Beurteilern in Präferenzexperimenten systematisch untersucht. Er weist darauf hin, daß die Auswahl nach interrater agreement die Annahme einschließt, alle Beurteiler bezögen sich auf die gleiche Variable, die nur in einer Dimension variere. Atypische Urteile würden auch dann nicht berücksichtigt, wenn sie reliabel sind (zur Gruppierung von Beobachtern nach der Ähnlichkeit ihrer Beurteilungsstrategien s. Naylor et al., 1967; Naylor & Schenck, 1966).

Wegen der besonderen Bedeutung, die der Verhaltensbeobachtung in der Verhaltenstherapie zukommt, ist im Arbeitskreis von O'Leary (z.B. Romanczyk et al., 1973) der *Schulung von Beobachtern*, meistens unter Reliabilitätsgesichtspunkten, größere Aufmerksamkeit zugewandt worden. Kontrollierte Übung führt fast immer zu einer bemerkenswerten Erhöhung der Übereinstimmung zwischen den Beobachtern. Nay & Kerkhoff (1974) zeigten für ein Kodiersystem mit 22 Symbolen, daß Feedback über Fehler die Reliabilität nach Abschluß des Trainings deutlich gegenüber einer Kontrollgruppe erhöht, die ohne Rückkoppelung am gleichen Videotape-Material lediglich kodieren übte. Training geschieht in jüngster Zeit häufig an Videotapeaufnahmen. Nay & Kerkhoff stellen folgende Vorteile heraus: 1. kann die Reliabilität in bezug auf eine standardisierte Vorlage abgeschätzt werden, 2. kann die Darbietung für Feedback an einen Beobachter unterbrochen werden, 3. können insbesondere schwierige Passagen beliebig oft, bis zur Beherrschung der Kodierung, wiederholt werden, 4. können in natura selten auftretende, jedoch wichtige Phänomene häufiger gezeigt werden; für ihr Auftreten kann sensibilisiert werden - generell läßt sich die Materialauswahl so vornehmen, daß alle Kategorien hinreichend oft geübt werden können.

Beobachter sind jedoch möglicherweise reaktiv: Ihre Einschätzungen werden verlässlicher, wenn man sie vorab darüber unterrichtet, daß Reliabilitätsprüfungen stattfinden. Wenn man Beobachtern Gelegenheit zum „Erfahrungsaustausch“ gibt, ändern sie im Verlauf der Beobachtungszeit ihre Interpretation der vorgegebenen Kategorien, nähern sie einander an und werden so reliabler, wenn auch nicht unbedingt valider (nach Romanczyk et al., 1973; s.a. Reid 1970, sowie Feger & Graumann in diesem Band).

3.4.3 Auswahl des zu beobachtenden Verhaltens

Mit Wright (1960, S. 73) kann man diese Auswahl unter inhaltlichem und unter zeitlichem Aspekt betrachten. „Material coverage . . . refers to what and how much at a time the observer tries to see in the stream of behavior.“ Hingegen bezieht sich „continuum coverage“ auf „... the length or parts into which the stream of behavior is divided for purposes of observation“. Diesen zweiten Aspekt bezeichnen wir als *Ziehen von Zeitstichproben* (time sampling). Longabaugh weist auf drei Entscheidungen hin, die beim Ziehen von Zeitstichproben zu fällen sind: 1. Ob man den Plan fixiert, vor Beginn der Beobachtung festlegt, oder offen läßt, wann wie lange etc. zu beobachten sei, und allenfalls Randbedingungen vorgibt, beispielsweise die, jede Person solle nicht öfter als einmal pro Tag beobachtet werden. 2. Wie lange der Zeitraum dauern soll, in dem das Ereignis beobachtet wird. 3. Wie lange das Pausenintervall zwischen Beobachtungen dauern soll. Die Gefahr beim Ziehen von Zeitstichproben liegt allgemein darin, die zeitliche Struktur des beobachteten Phänomens nicht angemessen zu erfassen. Andererseits wird kontinuierliche Beobachtung, z.B. eines gesamten Therapieverlaufs, nicht immer möglich sein (Literatur: Arrington 1939, 1943; Hutt & Hutt 1974; Wright 1960).

Beim Ziehen von Zeitstichproben kann man entweder ein einziges Intervall wählen und es festlegen, indem man Anfangs- und Endzeitpunkt bestimmt. Man kann sich auch für mehrere Intervalle entscheiden, dann sind deren Länge, u.U. auch die Pausenlänge und Startzeitpunkte zu vereinbaren. Ein wesentlicher Gesichtspunkt ergibt sich dabei aus der Art der Beobachtungen, die man erwartet, insbesondere aus deren *vermuteter Dauer und Häufigkeit*. Sackett (1978) stellt folgendes Schema auf (s. S. 19 oben).

Als Optimierungskriterium für die Ausschnittwahl formuliert Sackett (S. 26): „The sampling problems of observational research involve maximizing the chances of actually observing any of these four types of behavior in sequences that are representative of the typical behaviors of the subjects under study.“ Um Verhalten vom Typ II und IV zu erfassen, müssen relativ lange Zeitstichproben mit kontinuierlicher Beobachtung gezogen werden; bei Typ I-Verhalten können die Stichproben kürzer sein, bei Typ III-Verhalten können die Beobachtungen diskontinuierlich vorgenommen werden. Auch die *Auswer-*

	häufig	selten auftretend
Verhalten kurzer und nahezu konstanter Dauer (momentary behaviors)	Z.B. Lidschlag TYP I	Z.B. Niesen TYP II
Verhalten variabler Dauer (duration meaningful)	Z.B. Gesprächsbeitrag in Dyade Typ III	Z.B. Akt physischer Aggression TYP IV

tungsziele, z.B. die Art der beabsichtigten Vergleiche und der zu prüfenden Hypothesen, bestimmen die Ausschnittwahl mit. Wenn z.B. Aussagen über Abfolgen gleicher oder verschiedener Verhaltensweisen oder über Zusammenhänge von Umweltzuständen und Verhalten beabsichtigt sind, liegt in der Regel kontinuierliche Beobachtung in relativ langen Intervallen nahe. Wenn nur Frequenz, nicht aber Dauer interessiert, können die Zeitstichproben unterschiedlich lang sein, sonst muß entsprechend adjustiert werden (ausführlich Sackett, 1978). Wenn mehrere Individuen zusammen auftreten und die Beziehungen zwischen ihnen interessieren, stellt sich die Frage, ob ein Beobachter eine Person oder mehrere zugleich beobachten soll. Eine Möglichkeit, den Ausschnitt festzulegen, besteht darin, ein bestimmtes Individuum auszuwählen (*focal individual coding method*) und während einer bestimmten Beobachtungsperiode alle interessierenden Verhaltensweisen dieses Individuums zu registrieren, insbesondere auch, mit wem es interagiert. Damit diese Methode zu repräsentativen Verhaltensstichproben für das Individuum und die Gruppe führt, sollte 1) jede Person reihum zum Fokus der Beobachtung werden, 2) bei jeder Interaktion bei der das fokale Individuum beobachtet wird, sollte festgehalten werden, mit wem und in welcher Art es interagiert, 3) für jede im Mittelpunkt stehende Person sollte Verhalten in einer genügend großen Anzahl von Situationen erfaßt werden, und 4) die interessierenden Verhaltensweisen sollten über die Gelegenheiten hinweg relativ stabil sein. Sind diese Bedingungen nicht erfüllt, so läßt sich aus Beobachtungen einzelner Individuen das Interaktionssystem der Gruppe nicht unverfälscht konstruieren. Je nach Zielen und Ressourcen des Forschers bleibt dann die Möglichkeit, alle Individuen gleichzeitig zu beobachten, u.U. mit technischen Hilfsmitteln, die Zahl der Kategorien zu vermindern, die Zahl der Beobachter zu vergrößern und deren Beobachtungen zu synchronisieren (Sackett, 1978).

Stehen mehrere Beobachter zur Verfügung, so sollten sie zufällig oder systematisch auf die Vpn oder die Versuchsbedingungen verteilt werden. Eine feste

Zuordnung führt zu der Gefahr einer Konfundierung von Unterschieden zwischen Beobachtern einerseits, Beobachteten oder Bedingungen andererseits.

3.4.4 *Übergreifende Auswahlstrategien*

Wir bezeichnen Anweisungen zur Auswahl dann als übergreifend, wenn sie sich auf mehr als eine Facette beziehen, z.B. nicht nur auf die Auswahl von Personen, sondern auch auf Situationen. In dem folgenden, leicht abgeänderten Schema von Longabaugh (1980, S. 78ff.) in Tab. 1 können Dauer von Beobachtung und Pausenintervall bei jeder Strategie fixiert sein oder variieren. Die Strategien unterscheiden sich danach, ob die zu beobachtenden Personen, die aufzusuchenden oder herzustellenden Umwelten und das zu erfassende Verhalten fixiert sind oder variieren. Fixiert heißt hier: bei der Planung und vor Beginn der Beobachtung festgelegt. Bleibt die Bedingung variabel, so bestimmt das beobachtete Ereignis, auf welche Situationen, Personen und Verhaltensweisen der Beobachter stößt.

Zur Erläuterung des Schemas besprechen wir einige Strategien. Beim Ziehen von *Ereignisstichproben* (event sampling) wird ein vorbestimmtes Verhalten nur dann aufgezeichnet, wenn bestimmte Personen es in einer zuvor festgelegten Umgebung zeigen. Munroe (1973, hier nach Longabaugh 1980, S. 80) beobachtete bei bestimmten Kleinkindern in zeitlich und räumlich vorher fixierten Umgebungen, wenn die Kinder schrien: wie lange sie dann weinten, nach wieviel Sekunden eine Pflegeperson erschien, wer das war und wie sie sich verhielt, usw. Das Ziel dieser Studie bestand darin, auf Verteilungen bezogene Aussagen zu vergleichen, die durch Ziehen von Ereignisstichproben gegenüber dem Ziehen von Zeitstichproben gewonnen wurden.

Ein Beispiel für das Ziehen von *Personen-Umwelt-Stichproben* ist die Arbeit von Schoggen (1976, s. Longabaugh 1980, p. 80), in der Dreijährige aus drei verschiedenen sozialen Schichten (a) beim Essen und beim freien Spiel, (b) im oder nahe bei dem Elternhaus des Kindes, und (c) in Anwesenheit der Mutter beobachtet wurden. Waren a, b und c gegeben, wurden eine Reihe von Verhaltensweisen aufgezeichnet. Letzteres - mehrere Mengen im Bildbereich - ist im Gegensatz zum Ziehen von Ereignisstichproben wesentlich für diese Strategie, und das Beobachtungsergebnis kann dementsprechend nicht nur aus dem Vergleich über die sozialen Schichten bestehen, sondern neben Aussagen über Verteilungsformen auch solche über korrelative Zusammenhänge der Verhaltensvariablen enthalten. Selten wird die Strategie verwendet, lediglich *Personen-stichproben* zu fixieren. Eine bekannte Ausnahme ist Barkers (1951) „One boys day“, in der zahllose Verhaltensweisen eines bestimmten Jungen in all seinen Umwelten eines einzigen Tages aufgezeichnet wurden. Das Beobachtungsergebnis stellt ein Inventar von Verhaltensweisen, Umwelten und ihren Zusammenhängen dar, bisweilen als *specimen record* bezeichnet (s. Weick 1968, S. 416f.; Wright 1960, S. 86).

Tabelle 1: Übergreifende Auswahlstrategien nach Longabaugh

Bezeichnung der Strategie	Personen	Umgebung	Verbalten
(1) Ziehen von Ereignisstichproben	fixiert	fixiert	fixiert
(2) Ziehen von Personen-Umwelt-Stichproben	fixiert	fixiert	variabel
(3) Ziehen von Personenstichproben	fixiert	variabel	variabel
(4) Ziehen von Umwelt-Verhaltens-Stichproben	variabel	fixiert	fixiert
(5) Ziehen von Umwelt-Stichproben	variabel	fixiert	variabel
(6) Ziehen von Verhaltensstichproben	variabel	variabel	fixiert
(7) Nichtfixiertes Ziehen von Stichproben	variabel	variabel	variabel

4. Die Bewertung von Beobachtungen

Damit Beobachtungen als wissenschaftlich gelten können, müssen sie bestimmte Kriterien notwendigerweise erfüllen. Die Begründung der Kriterien als notwendige, m.E. ein Thema der Wissenschaftstheorie, kann hier nur angedeutet werden. Die Planung von Beobachtungen geschieht auch mit dem Ziel, in der erwarteten Bewertung diesen notwendigen Kriterien zu genügen, wobei in der Vorbereitungsphase oft Planungsschritte und solche der Bewertung rückgekoppelt werden, bis die Kriterien hinreichend erfüllt sind. Von den notwendigen unterscheiden wir differenzierende Kriterien. Dabei gehen wir davon aus, daß alle empirischen Methoden in Psychologie und Sozialwissenschaften als Varianten der wissenschaftlichen Beobachtung aufgefaßt werden können, wobei die wesentlichen Unterschiede zwischen den Methoden sich aus dem Zweck ergeben, dem sie dienen sollen. So ist der Zweck des Tests, inter- und intraindividuelle Unterschiede zu bestimmen, Strukturanalysen wie multidimensionale Skalierung oder Clusteranalysen haben die Aufgabe, Ordnungsstrukturen nachzuweisen, und das Experiment sucht nach Zusammenhängen zwischen Variiertem und Variierendem. Wegen dieser Absicht werden die Bedingungen, unter denen die Variablen des Experiments manipuliert und erfaßt werden, im idealen Experiment so kontrolliert, daß eindeutige Schlüsse von Antezedenzbedingungen auf Beobachtungsvariablen möglich sind. Bei den meisten Tests sind Durchführungssituation und Reaktionsmöglichkeiten in der Form der Standardisierung kontrolliert, wieder mit dem Ziel, Beobachtungen im interindividuellen Vergleich möglichst eindeutig und zutreffend deuten zu können. Test und Experiment sind also nicht, wie bisweilen behauptet wird, Alternativen zu oder Konkurrenten der wissenschaftlichen Beobachtung, sondern Beobachtungen unter zweckspezifischen Kontrollmaßnahmen. Beobachtungsstudien i.e. S. werden oft dann durchgeführt, wenn aus verschiedenartigen Gründen diese Kontrollen nicht durchgeführt werden können, z.B. wenn man die unabhängigen Variablen nicht variieren kann oder darf. Oft auch leisten Beobachtungsstudien die Pionierarbeit, das Universum der Beobachtungen einzugrenzen und zu strukturieren.

Während der Grad der Kontrolle variieren kann und die Art der Kontrolle dem jeweiligen Ziel angepaßt ist, stellt Kontrolliertheit wieder nur Mittel zum Zweck dar: verlässliche und gültige Schlüsse ziehen zu können. Reliabilität und Validität sind jene notwendigen Kriterien, die in der Psychologie das meiste Interesse gefunden haben. Reliabilität spielt als Kriterium in der Psychologie deshalb eine so große Rolle, weil wir oft nicht wissen, welche Variablen wie zu kontrollieren sind, oder diese Kontrollen nicht anwenden können, um jene Bedingungen herzustellen, deren Realisation die Theorie verlangt, und jene Variationsquellen auszuschalten, mit denen sich die jeweilige Theorie nicht befaßt und die insofern gerade irrelevant sind. Reliabilität kann daher allgemein als das Verhältnis von systematischer, d.h. theoretisch erklärter Varianz,

zu Fehlervarianz, d.h. durch die jeweils betrachteten Bedingungen nicht erklärte Varianz, konzipiert werden. Wenn die theoretischen Ideen, die den Abbildungssatz formulierten, zutreffen, dann ist es eine Frage der Kontrolle und ein Ausdruck der Reliabilität, wie stark der empirische Zusammenhang zwischen Definitionsbereich und Bildbereich des Abbildungssatzes ist.

Kategoriensysteme unterscheiden sich, wie erwähnt, nach dem Ausmaß, in dem Beobachtetes interpretiert werden muß, damit es der Beobachter kodieren kann. Longabaugh (1980, S. 102) zeigt, welche Konsequenzen dies für die Übereinstimmung von Beurteilern hat, wenn sie auf den ihnen gemeinsamen kulturellen Hintergrund angewiesen sind, um Verhalten eindeutig zu interpretieren:

„Even within the same culture persons have only partial access to shared meaning and an imperfect understanding of what they have access to. When several observers are asked to categorize a particular unit of behavior, the greater the number of observers, the less the average agreement obtained. . . . As the number of observers increases, the meaning shared by all remains at best constant. Meanwhile the component of meaning shared by fewer than all observers increases as does the idiosyncratic meaning unique to each observer.“

Nicht nur die Verlässlichkeit, sondern auch die Gültigkeit von Urteilen, die auf Verhaltensbeobachtung basieren, hängt von dem gemeinsamen Verständnis der sozialen und kulturellen Umwelt ab. Verständlicherweise wird dies besonders in kulturvergleichenden Untersuchungen deutlich, weshalb wir noch einmal Longabaugh (1980, S. 102) zitieren können:

„Veridical judgments regarding actor intention require that *actor* and *observer* share the same symbols; veridical judgments regarding the effect of the behavior require that *observer* and *target* share the same symbols; veridical judgments concerning the meaning of the behavior for the relationship require that *actor*, *target*, and *observer* share the same symbol. Veridical judgment concerning the cultural significance of the behavior requires that *observer*, *actor*, *target*, and *most* other *cultural participants* (in principle) agree upon the meaning of the symbol.“

5. Die Reproduzierbarkeit von Beobachtungen

Reliabilität hat je nach der mit Beobachtungen konfrontierten Theorie verschiedene Aspekte. Sagt die Theorie nichts über Unterschiede zu verschiedenen Meßzeitpunkten, dann läßt sich die Kontrolliertheit unter dem Gesichtspunkt der zeitlichen Stabilität problematisieren. Spezifiziert die Theorie keine Effekte unterschiedlicher Meßinstrumente, so sollten die Ergebnisse mit Hilfe unterschiedlicher Instrumente wiederholbar sein. Generell beschreibt also Re-

liabilität die Reproduzierbarkeit von Beobachtungen unter theoretisch für das Auftreten des Beobachteten äquivalenten Bedingungen bei Unterschieden in theoretisch irrelevanten Bedingungen. Was als wesentliche oder als irrelevante Bedingungen anzusehen ist muß von der Fragestellung der Untersuchung her begründet werden. Eine Reliabilitätsstudie legt fest, was erfaßt werden soll, welche Bedingungen konstant sind und welche variabel; letztere stellen die Hinsicht dar, in der die Ergebnisse reproduzierbar sein sollen - das können z.B. Meßwiederholungen zu verschiedenen Zeitpunkten, mit verschiedenen Apparaten und durch verschiedene Personen sein. Formal unterscheiden wir zwischen (1) den verschiedenen Meßinstrumenten, für die hier interessierende Reliabilitätsforschung sind das durchweg *Beobachter*, (2) den verschiedenen *Beobachtungsgegenständen*, hier meist verschiedene Personen, (3) den verschiedenen *Beobachtungshinsichten*, hier in der Regel als inhaltlich verschiedene Kategorien konzipiert, die an der gleichen Person erfaßt werden können, und schließlich (4) verschiedenen *Beobachtungsebenen*, oft in Form von wiederholten Messungen realisiert. Die Reproduzierbarkeit von Beobachtungsergebnissen studiert man über die Variation wenigstens einer dieser vier Mengen; oft interessiert die Vergleichbarkeit der Ergebnisse, die mit verschiedenen Meßinstrumenten - durch diverse Beobachter - gewonnen wurden, oder über unterschiedliche Meßzeitpunkte. Die Reliabilitätsarten der Testtheorie (s. Kap. von Kristof und Fischer in diesem Band) finden hier ihre logische Entsprechung.

Jede Beobachtung und Messung, auch die im Bereich der klassischen Naturwissenschaften vorgenommene, ist mit mehr oder weniger großen Meßfehlern behaftet. Dafür kann man sich zahlreiche Gründe vorstellen, von der falschen Justierung des Meßgerätes bis zur Unaufmerksamkeit des Ablesenden. Da man mit solchen Fehlern rechnet, interessieren die Fragen, wie groß der Fehler ist, ob er konstant, systematisch ist (= bias) oder ob er zufallsmäßig variiert (= error). Häufig geht man davon aus, man sei in der Lage, größere, systematische Fehler zu identifizieren, dann auszuschalten oder zu kontrollieren - meist bei der Konstruktion des Meßgerätes oder der Formulierung von Kategorien und Instruktion. Als Beispiel für diesen Prozeß in der Psychologie sei die Arbeit von Hendel & Weiss (1970) erwähnt, in der ein Nachweis *stabiler* interindividueller Unterschiede in der Konsistenz bei Paarvergleichen gelang. Bestimmte Reliabilitätseinschätzungen können also erhöht werden, wenn die konsistenteren Beurteiler ausgewählt oder die Urteile z.B. durch Schulung konsistenter werden. Was bleibt, sind dann zahlreiche, meist unidentifizierte Quellen, die kleine Zufallsfehler bewirken, über deren Wirkungen man annimmt, daß sie sich additiv kombinieren und einen Erwartungswert von Null haben. Also muß zur Fehlerabschätzung wie zur Bestimmung des richtigen, des „wahren Wertes“ dieser Grenzwert von Messungen approximiert werden, was innerhalb der Naturwissenschaften häufiger getan wird und zu Serien von Hunderten von Meßwiederholungen führt. Als Fehlermaß wird dann die mitt-

lere Abweichung (average error) bestimmt, wie aus der folgenden Tabelle hervorgeht:

Messung	Ergebnis	absolute Abweichung vom Mittelwert in mm
1	100 m m	0
2	101 m m	1
3	104 m m	4
4	99 m m	1
5	96 m m	4
$\Sigma 10; n = 5 \text{ Messungen}$		

mittlere Abweichung = $\bar{x} = 2 \text{ mm}$

Zur besseren Vergleichbarkeit über verschiedene Skalen wird auch der prozentuale Fehler (percentage error) angegeben, der als mittlere Abweichung relativ zur Größe des Objekts definiert ist. Betrachten wir im Beispiel 100 mm als Objektgröße, dann beläuft sich der prozentuale Fehler auf $\frac{2}{100} = 2\%$.

Messungen und damit Reliabilitätsprobleme wurden in der Psychologie verständlicherweise dann besonders brisant, als von Testwerten existenzielle Entscheidungen abhingen, so daß im Bereich der Testtheorie und Testkonstruktion eine lange Tradition der Reliabilitätsforschung zu finden ist. Ohne hier auf die zahlreichen Gründe dafür einzugehen, darf man allerdings davon ausgehen, daß eine Person nur selten tausendmal -wie ein Objekt der Naturwissenschaften - gemessen werden kann. Die Lösung dieses Problems in der klassischen Testtheorie besteht darin, statt vieler Messungen an einer Vp zwei oder wenig mehr Erhebungen an zahlreichen Vpn durchzuführen. Um diesen Ausweg beschreiten zu können, muß man annehmen, man könne Zufallsstichproben aus der für eine Messung in Frage kommenden Population von Vpn ziehen und dabei stehe für diese Meßzwecke jede Vp im Prinzip stellvertretend für jede andere. Als Fehlermaß bietet sich analog zum mittleren Fehler die durchschnittliche absolute Abweichung von erster und zweiter Messung an. Dieses Maß wird jedoch in der Psychologie selten angewandt, statt dessen wird ein Produkt-Moment-Korrelationskoeffizient bestimmt, wodurch mögliche Unterschiede zwischen den Mittelwerten und Streuungen beider Meßwertreihen sich nicht auf das Fehlermaß auswirken können. Dieses Vorgehen kommt der Annahme gleich, Differenzen in Skaleneinheit und Skalensprung seien irrelevant, vielmehr seien die relativen Unterschiede zwischen den Personen wesentlich.

Die andere Tradition der Reliabilitätsforschung ist mit der Methodenlehre der wissenschaftlichen Beobachtung verbunden. Hier tritt an die Stelle des Meßinstrumentes oder Tests der Beobachter als Datenquelle, und erhoben wird die Reproduzierbarkeit seiner Aussagen über wiederholte Gelegenheiten als *intra-rater-consistency* und die Vergleichbarkeit der Befunde verschiedener Beobachter als *inter-rater-agreement*. Die Methodik der Reliabilitätsprüfung hat in den letzten Jahrzehnten auf der Auswertungsseite erhebliche Fortschritte gemacht, die wir im folgenden skizzieren. Die veröffentlichten Beobachtungsstudien haben diese Entwicklung oftmals noch nicht rezipiert. Susman et al. (1976, nach Hollenbeck, 1978) untersuchten Beobachtungsstudien in 15 Zeitschriften aus entwicklungspsychologischen, klinischen und pädagogischen Bereichen und stellten fest: 32% der Arbeiten berichteten überhaupt keine Reliabilitätsangaben. Fast alle Schätzungen, die referiert wurden, waren Angaben über prozentuale Übereinstimmung (s.u.), und während der 16 Jahre, die sie diese Zeitschriften verfolgten, fanden die Autoren keine Verbesserung.

Um die große Zahl der Analyseverfahren zu ordnen, *gliedern wir nach dem Skalenniveau*, das den Beobachtungen als Daten zugeschrieben wird. Denn man wird Reproduzierbarkeit nur hinsichtlich jener Datenmerkmale fordern, die man in den Daten als vorhanden annimmt; wenn also nur Rangskaleninformation in den Daten vorhanden ist, wird man nicht auf Übereinstimmung in den Intervallgrößen prüfen (allgemeine Übersichtsliteratur: Asendorpf & Wallbott, 1979; Frick & Semmel, 1978; Landis & Koch, 1975; Lienert, 1973, Kap. 9; Tinsley & Weiss, 1975; Computerprogramme z.B. Cicchetti et al., 1977).

5.1 Übereinstimmungsmaße für nominalskalierte Daten

5.1.1 Prozentuale Übereinstimmung und allgemeine Vorüberlegungen

Von Übereinstimmung sprechen wir, wenn das Beobachtete identischen Kategorien zugeordnet wird.

Zu den in der Literatur sehr häufig berichteten und in der Regel auf dichotome Daten angewandten Übereinstimmungsmaßen gehört der *Prozentsatz*, mit dem (zwei) Beobachter das gleiche Material in die gleichen Kategorien ordnen. Dabei nimmt man implizit an: Die Übereinstimmung bei der Beurteilung des einen Beobachtungsgegenstandes ist genauso (positiv) zu werden wie die Übereinstimmung bei jedem beliebigen anderen und die abweichende Klassifikation bei dem einen Material ist ein gleichschwerer Fehler wie jede andere Nichtübereinstimmung. Unterscheidet sich dann die Materialstichprobe bei der Erprobung des Kategoriensystems in ihrer Klassifikationsschwierigkeit von der Materialstichprobe der eigentlichen Erhebung (z.B. Videotraining vs.

Feldstudie), dann kann man nicht gleiche Reliabilitätskoeffizienten erwarten. Dies gilt auch für den Austausch von Beobachtern, wenn man Unterschiede in der Klassifikationsgüte erwarten muß. Sofern Reliabilitätsstudien weder Material- noch Beobachterparameter berücksichtigen, sind die Grenzen der Verallgemeinerung der Reproduzierbarkeitsangaben unklar.

Tabelle 2: Daten nach Hollenbeck (1978).

Beobachter	Kategorie	Zeitblöcke										
1		1	2	3	4	5	6	7	8	9	10	Σ
	A	0	1	1	0	0	1	0	0	1	1	5
	B	1	0	1	0	0	0	1	0	0	0	3
	C	0	1	1	0	1	0	1	0	1	0	5
2												
	A	0	1	0	0	1	1	1	0	1	1	6
	B	1	0	0	1	0	1	0	0	0	0	3
	C	1	1	1	0	1	0	0	0	1	0	5

Legende: 1 = Verhalten in der entsprechenden Kategorie wurde beobachtet,
0 = sonst.

Hollenbeck (1978) diskutiert die Problematik von Übereinstimmungsprozent-sätzen an den Beobachtungen in Tab. 2, wobei für „Zeitblöcke“ jede Art von Wiederholung stehen kann, beispielsweise auch „Vpn“ oder „Situationen“. - Zunächst ergeben sich nun mehrere Möglichkeiten, prozentuale Übereinstimmung (% \ddot{U}) zu berechnen, z.B. für jede Kategorie einzeln, für den Durchschnitt, und für vollständige Übereinstimmung pro Zeitblock:

Kategorie	Anzahl konkordanter Paare	% \ddot{U}
A	7 von 10	70
B	6 von 10	60
C	8 von 10	80
Durchschnitt	7 von 10	70
vollständig pro Zeitblock	4 von 10	40

Weitere Möglichkeiten entstehen, wenn man Gewichte für verschiedene Arten von Konkordanz oder Diskordanz einführt (z.B. Übereinstimmung bei positiver Identifikation des Verhaltens sei gewichtiger zu veranschlagen als bei negativer, s.u.), oder statt wie im Beispiel 0/1 zu registrieren, feststellt, wie oft das Verhalten im jeweiligen Zeitblock auftrat, oder die Sequenz berücksichtigt, in der Verhaltensweisen innerhalb eines Zeitintervalles beobachtet wurden.

Die Kritik an % \bar{U} faßt Hollenbeck in folgenden Punkten zusammen: 1. Die Höhe des möglichen % \bar{U} hängt von den Randsummen (letzte Spalte in Tab. 2) ab. Je unterschiedlicher die Randsummen sind, desto geringer ist die maximal mögliche Übereinstimmung. Dies ist solange wünschenswert, wie sich in unterschiedlichen Randsummen ausschließlich Fehler der Beobachter niederschlagen können, etwa zu geringe Sensibilität jener Beobachter mit niedrigen positiven Randsummen. Wie jedoch Randsummen-Differenzen zu deuten sind, läßt sich nur aus der jeweiligen Studie bestimmen, und eine Anwendung der signal detection theory (Green & Swets, 1966, einführend: McNicol, 1972) wäre möglich, wenn man Sensitivitätsunterschiede und unterschiedliche Reaktionstendenzen prüfen möchte 2. Die *Bewertung* des berechneten % \bar{U} hängt von den Randsummen ab, denn aus ihnen läßt sich bestimmen, wie groß die rein zufällig zu erwartende Übereinstimmung ist (s.u.). Für Kategorie A in Tab. 2 ergibt sich als *Erwartung* aufgrund der Randsummen:

Beobachter 2

Beobachter 1

	0	1	
0	2	3	5
1	2	3	5
	4	6	10

erwartet

3	2
1	4

beobachtet

Also sind $\frac{4 \times 5}{10} + \frac{5 \times 6}{10} = 5$ Übereinstimmungen aufgrund der Randsummen zu erwarten, somit ist die beobachtete prozentuale Übereinstimmung von 70% genau 20% besser als die Zufallserwartung von 50%, bei Kategorie B ist sie nur 2% besser, bei C 30% besser als der Zufall. 3. % \bar{U} informiert nicht darüber, *wo die Fehler liegen* und welcher Art sie sind. 4. Das durchschnittliche % \bar{U} über alle Kategorien kann - besonders, wenn es nicht extrem hoch ist und über viele Kategorien zusammengefaßt wird - drastische *Unterschiede zwischen den einzelnen Kategorien* verdecken. Wenn inhaltliche Schlüsse auf Beobachtungen in einzelnen Kategorien basieren, muß die Verlässlichkeit für

diese Kategorien bekannt sein. 5. Wenn die Gelegenheiten zur Wiederholung der Beobachtungen als Zeitintervalle geplant werden, variiert % Ü als **Funktion der Länge dieses Intervalles**: Je länger, desto größer die vielleicht nur scheinbare Übereinstimmung, wenn nur das bloße - mindestens einmalige - Auftreten des Verhaltens registriert wird.

Zu den wichtigsten Einsichten der Reliabilitätsmethodik gehört die Erkenntnis, daß Übereinstimmung nicht durch jedes Zusammenhangs- oder Assoziationsmaß beschrieben werden kann. Das allgemeine statistische Problem, den Zusammenhang zwischen zwei oder mehr nominalskalierten Variablen zu messen, wurde ausführlich von Goodman & Kruskal (1954, 1959, 1963, 1972) erörtert (Diskussion dieser Arbeiten bei Bishop et al. 1974, McKinlay, 1975). Übereinstimmung ist jedoch ein Spezialfall von Zusammenhang, wie aus folgendem Beispiel hervorgeht:

		1. Beobachter		
		A	B	C
2. Beobachter	Kategorie A	0	50	0
	Kategorie B	0	0	50
	Kategorie C	50	0	0

Der Zusammenhang ist perfekt, die Übereinstimmung jedoch gleich Null. Damit Übereinstimmung vorliegt, müssen die Fälle in identische Kategorien geordnet werden, für Zusammenhang genügt eine solche Zuordnung, welche die Vorhersage der Klassifikation durch den einen Beurteiler aus der Klassifikation eines anderen Beurteilers gestattet.

5.1.2 Systematik einiger Übereinstimmungsmaße **für** nominalskalierte Daten

Eine weitere, für die Reliabilitätsmethodik wesentliche Einsicht besteht in der Erkenntnis, daß Übereinstimmung sich auch zufällig einstellen kann, wenn Beobachter gar nicht beobachten und so gewonnene Information zur Grundlage ihres Urteils machen, sondern blind - z.B. durch Münzwurf - kategorisieren, oder wenn sie aufgrund verschiedener, voneinander unabhängiger Kriterien zu ihren Resultaten kommen. Wenn jeder von zwei Beurteilern jede von zwei Kategorien gleich oft verwendet, ist zu erwarten, daß sie durchschnittlich in der Hälfte der Fälle übereinstimmen werden. *Die beobachtete Übereinstimmung sollte also in irgendeiner Form gegenüber der zufällig zu erwartenden Übereinstimmung relativiert werden.* Schutz (1952) war in dieser Forschungs-

richtung wohl der erste, der ein Modell für den Urteilsprozeß mit einer Zufallskomponente ausdrücklich formulierte. Er berechnet den Übereinstimmungsprozentsatz, der beobachtet werden muß, damit der Forscher auf einem vom ihm gewählten Konfidenzniveau davon ausgehen kann, daß ein von ihm als mindestens notwendig erachtetes Übereinstimmungsniveau, das frei von Zufallsübereinstimmung ist, tatsächlich erreicht wurde.

Das Zufallsmodell, das Schutz seiner Korrektur zugrunde legte, ist etwas kompliziert und trifft wenig plausible Annahmen, z.B. daß kein Beobachter, wenn er im Zustand der zufälligen Beurteilung ist, eine Präferenz für eine der Kategorien hat. Scott (1955) legte m. W. als erster einen Koeffizienten π vor, in dem die Zufallskorrektur direkt vorgenommen wird, und zwar als Subtraktion der zufälligen Übereinstimmung P_e von der beobachteten Übereinstimmung P_o . Die Differenz $P_o - P_e$ wird dann standardisiert mit Hilfe einer Division durch $1 - P_e$ (das ist der maximal mögliche Wert, den die Differenz annehmen kann), wodurch erreicht wird, daß π bei vollständiger Übereinstimmung = + 1; bei Übereinstimmung, die nicht über die Zufallserwartung hinausgeht, ist $\pi = 0$. Somit

$$\pi = \frac{P_o - P_e}{1 - P_e}$$

Für den Fall von 2 Beobachtern und einer dichotomen Kategorie führen wir folgende Notation ein:

		2. Beobachter		
		1	0	
1. Beobachter	1	a	b	p_1
	0	c	d	q_1
		p_2	q_2	1

a, b, c, d sind die relativen Häufigkeiten in den Zellen, die Randsummen sind ebenfalls als Proportionen dar gestellt. P_o wird als % Ü, also = $a + d$ berechnet. P_e wird berechnet als $\left(\frac{p_1 + p_2}{2}\right)^2 + \left(\frac{q_1 + q_2}{2}\right)^2$, was inhaltlich dem Versuch entspricht, die zufällige Übereinstimmung auf der Basis von Populations-Randwahrscheinlichkeiten und der Annahme zu berechnen, beide Beobachter wiesen die gleiche Randverteilung auf, die der in der Population entspräche (Light, 1971, S. 367, zur Kritik an Scott s. auch Lisch & Kriz 1978).

Cohen (1960) führte einen inzwischen sehr bekannt gewordenen Koeffizienten κ ein, der sich - abweichend von π - auf die beobachteten Randverteilungen bezieht und nicht davon ausgeht, die Randverteilungen seien gleich. Das Rationale für die Zufallskorrektur ist also das gleiche wie bei π , jedoch berechnet

sich hier P_e als $p_1p_2 + q_1q_2$. Light (1971) erweiterte κ , und zwar sowohl für mehr als zwei Beobachter, ein „overall group agreement measure“ (neuere Übersicht hierzu bei Conger 1980) als auch für ein zu den Randverteilungen konditionales übereinstimmungsmaß (Programm: McDermott & Watkins, 1979). Ferner berichtet er einen Signifikanztest für die gemeinsame Übereinstimmung mehrerer Beobachter, wenn sie mit einem Standard - z.B. einer „richtigen“ Klassifikation - oder wenn ein bestimmter Beobachter mit allen anderen verglichen werden sollen (Programm: Watkins & McDermott, 1979). Wackerly et al. (1978) haben diese Thematik für den Fall des Vergleichens eines Beobachters mit einem bekannten Standard weiterentwickelt. Sie unterscheiden zwei Fälle: Dem Beobachter wird die Randverteilung vorgegeben, oder er kann sie selbst bestimmen. Im ersten Fall bestehen Abhängigkeiten zwischen den Urteilen, die man für die Berechnung der Zufallsübereinstimmung berücksichtigen muß. Wackerly et al. berichten eine Möglichkeit zur inferenzstatistischen Prüfung von κ auf Überzufälligkeit auch unter der Annahme intraindividuelle Abhängigkeit der Urteile eines Beobachters.

Um die *Rolle der Randverteilungen* für die Wahl eines Übereinstimmungsmaßes zu verdeutlichen, vergleichen wir zwei Fälle:

Fall I

1. Beobachter

1 0

1	60	20	80
0	20	0	20

80 20

2. Beobachter

Fall II

1. Beobachter

1 0

1	20	60	80
0	0	20	20

20 80

Im ersten Fall scheint die Übereinstimmung mit 60% \ddot{U} größer zu sein als in Fall II mit 40% \ddot{U} . Geht man jedoch von den Restriktionen durch die vorgegebenen Randsummen aus, so zeigt sich in I die geringstmögliche, in II die größtmögliche Übereinstimmung. Die deutliche Asymmetrie in der Kategorienbenutzung schafft Probleme für die Interpretation der Übereinstimmung; Light nennt % \ddot{U} die absolute, die randverteilungskorrigierte Übereinstimmung die relative. Wieder hängt es von den Umständen in der jeweiligen Untersuchung ab, ob man in das Übereinstimmungsmaß eine Korrektur für abweichende Randverteilungen einbeziehen soll.

Als Erweiterung hat Cohen (1968) das gewichtete κ eingeführt, das es erlaubt, Nichtübereinstimmung nach der Schwere der Folgen zu gewichten, wobei der Forscher die Gewichte festlegen muß.

Fleiss (1971) verallgemeinerte κ für die Situation, daß jeder Beobachtungsfall auf Nominalskalenniveau von der gleichen Anzahl Beurteiler eingestuft wird, aber die Beurteiler, die den einen Fall beobachten, nicht notwendigerweise die gleichen sind, die einen anderen Fall beobachten. Fleiss schildert auch die Möglichkeit zu bestimmen, wie hoch die Übereinstimmung darüber ist, eine bestimmte Person oder Verhaltensweise einer bestimmten Kategorie zuzuordnen. Fleiss et al., 1972, behandeln den Fall, daß eine Person durch mehr als eine Variable beschrieben wird. - Wenn bei der Berechnung der Übereinstimmung die zeitliche Abfolge berücksichtigt werden soll, also Sequenz und jeweilige Dauer der beobachteten Ereignisse, ergeben sich besondere Schwierigkeiten, auf die Hollenbeck (1978) eingeht, der κ auch in diesem Fall anwendet (s.a. Asendorpf & Wallbott, 1979, sowie Abschn. 5.4). - Um prüfstatistische Fragen behandeln zu können, wurde für κ und gewichtetes κ eine Stichprobentheorie entwickelt (Everitt 1968, Fleiss et al., 1969, Übersicht bei Hubert 1977, für kleine Stichproben Wackerly et al., 1978, man beachte die korrigierten Formeln 'in Fleiss et al., 1979). Die inferenzstatistischen Ansätze gehen durchweg davon aus, daß die wiederholten Beobachtungen eines Beobachters oder verschiedener Beobachter voneinander unabhängig sind. Diese Annahme ist prinzipiell empirisch prüfbar, beispielsweise durch den Vergleich der Übereinstimmung zwischen einzelnen arbeitenden Beobachtern und solchen, die schon dadurch interagieren, daß sie die Beobachtung im gleichen Raum durchführen. Bisweilen führt schon die Planung der Erhebung dazu, daß die aus statistischen Gründen geforderte Unabhängigkeit der Beobachtungen kaum zu erwarten ist, beispielsweise, wenn den Beobachtern gestattet wird, ihre Beurteilung des früher auftretenden X nach Beobachtung des später auftretenden Y zu ändern. Unabhängigkeit der Beobachtungen kann auch bedeuten: Die Ergebnisse der Beobachtungen bei Wiederholungen sind nicht eine Funktion der Wiederholungen. Das erleichtert auch die inhaltliche Interpretation der Befunde, denn im Idealfall ändert wiederholte Beobachtung weder das Beobachtete noch das Beobachtungsinstrument.

Die meisten Übereinstimmungsmaße gehen davon aus, daß die Beobachter die gleichen, vor der Beobachtung definierten Kategorien benutzen. Man kann es jedoch auch den Beobachtern freistellen, welche und wieviele Kategorien sie definieren. Beispiele wären: Beurteiler verwenden verschiedene nosologische Klassifikationen, Lehrer verschiedene Notensysteme, Vpn beschreiben ihr Lösen der gleichen Denkprobleme mit unterschiedlichen Begriffen. Dann muß man dennoch nach Brennan & Light (1974) nicht darauf verzichten, Übereinstimmung zu berechnen. Die Idee besteht darin, Paare von Beobachtungsergebnisse zu betrachten: Ordnen von zwei Beobachtern der eine zwei Beobachtungen in die Kategorie „flüssiger Verhaltensablauf“, der andere beide Beobachtungen in seine Kategorie „geschickter Bewegungsvollzug“, so erfassen beide Kategorien Gleiches trotz inhaltlich verschiedener Bezeichnungen, und eine übereinstimmende Klassifikation eines Paares in dieselbe Kategorie stellt

einen Hinweis auf Übereinstimmung der Beobachter dar wie auch der Fall, daß beide die Elemente eines Paares von Beobachtungen in jeweils andere Kategorien einordnen. Hubert (1977a) leitet Mittelwert und Varianz des Koeffizienten von Brennan & Light ab und ermöglicht so die Prüfung auf signifikante Abweichung von Null und die Bestimmung von Konfidenzintervallen. Hubert zeigt auch, wie dieser Koeffizient auf den Fall geordneter Kategorien ausgedehnt werden kann.

Es gibt inzwischen eine große Zahl von Übereinstimmungskoeffizienten, von denen hier noch der G-Index erwähnt werden soll, den Holley & Guilford (1964) zunächst als Maß der Ähnlichkeit von zwei Personen, charakterisiert über n dichotome Items eingeführt haben. Holley & Lienert (1974) beschreiben ihn als Übereinstimmungsmaß mit einer typischen Anwendung, daß zwei oder mehr Beobachter eine oder mehrere Personen hinsichtlich n dichotomer Merkmale charakterisieren. Weitere Verallgemeinerungen finden sich bei Vegelius (1977, 1977a, 1979) und in der dort referierten Literatur. - Die große Zahl von übereinstimmungsmaßen hat zu statistischen Arbeiten geführt, die die Koeffizienten vergleichen und ihre impliziten Annahmen systematisieren (z.B. Fleiss & Cohen, 1973; Janson & Vegelius, 1979; Hubert, 1979b). Krippendorff (1970a) geht von folgender, häufig verwendeter Formel für Übereinstimmungskoeffizienten aus:

$$(1) \quad \text{Übereinstimmung} = 1 - \frac{\text{beobachtete Nichtübereinst.}}{\text{erwartete Nichtübereinst.}}$$

Der Koeffizient wird 0, wenn die Übereinstimmung rein zufällig ist, 1 wenn sie vollkommen ist, negativ, wenn sie hinter der Zufallserwartung zurückbleibt. Da nur Übereinstimmung, nicht Zusammenhang allgemein interessiert, müssen die Zellen einer Frequenzmatrix für die Beobachter i und j bei der Berechnung der Nichtübereinstimmung gewichtet werden mit d_{ij} , wobei für ungeordnete Kategorien gilt:

$$(2) \quad d_{ij} = \begin{cases} 0, & \text{wenn } i = j, \\ 1, & \text{wenn } i \neq j. \end{cases}$$

Wenn die Kategorien geordnet sind oder durch Skalenwerte mit Intervallskalenniveau repräsentiert werden können (s.u.), wird folgende Gewichtungsfunktion vorgeschlagen:

$$(3) \quad d_{ij} = (i - j)^2, \text{ mit } i, j \text{ als Kennwerten der Kategorien.}$$

Die erwarteten Frequenzen (e_{ij}) werden entweder bestimmt unter Annahme gleicher Randverteilungen:

$$(4) \quad e_{ij} = \frac{1}{n} \left(\frac{n_{i.} + n_{.j}}{2} \right) \left(\frac{n_{i.} + n_{.j}}{2} \right)$$

wobei die folgende Frequenzmatrix die Notation definiert:

		2. Beobachter	
		Kategorie	j
1. Beobachter	i		
		n_{ij}	$n_{i.}$
		$n_{.j}$	$n = \text{Gesamtzahl der beurteilten Einheiten}$

Dabei bedeutet n_{ij} die Anzahl der Fälle in Zeile i und Spalte j. Ein Punkt im Index zeigt an, daß über die entsprechende Zeile oder Spalte summiert wurde.

Unter Berücksichtigung ungleicher Randverteilungen ergibt sich:

$$(5) \quad e_{ij} = \frac{1}{n} n_{i.} n_{.j} .$$

Wenn als Maß für die beobachtete Nichtübereinstimmung

$$\sum \sum n_{ij} d_{ij}$$

definiert wird und für die zufallsbedingte Nichtübereinstimmung

$$\sum \sum e_{ij} d_{ij} ,$$

so erhält (1) die Form:

$$(6) \quad a = 1 - \frac{\sum \sum n_{ij} d_{ij}}{\sum \sum e_{ij} d_{ij}} .$$

Scotts π erhält man durch Einsetzen von (2) und (4) in (6), Cohens κ durch Einsetzen von (2) und (5) in (6), den Intraklassen-Korrelationskoeffizienten nach Pearson (1901) durch Einsetzen von (3) und (4) in (6) und Spearmans Rangkorrelationskoeffizienten ρ durch Einsetzen von (3) und (5) in (6). Nach Krippendorffs Systematik hat der Forscher also zwei Auswahlentscheidungen zu treffen, ob Unterschiede in den Randverteilungen zum Fehler geschlagen werden sollen oder nicht, und ob die Kategorien eine Ordnung aufweisen oder nicht. In der systematischen Übersicht von Fleiss (1975) kommt als weiterer

Klassifikationsgesichtspunkt die Frage hinzu, ob ein Übereinstimmungsmaß die Proportionen der Übereinstimmung über das Vorliegen eines Merkmals und der Übereinstimmung über Nichtvorliegen symmetrisch behandeln oder nicht. Wenn, wie bei taxonomischen Fragestellungen nicht selten, positive Merkmalsausprägung sehr selten, Fehlen des Merkmals jedoch häufig auftritt, kann es sinnvoll sein, in einem Übereinstimmungsmaß nur die Fälle zu berücksichtigen, in denen von wenigstens einem Beobachter positive Merkmalsausprägung festgestellt wurde (so z.B. Dice, 1945; weiteres bei Fleiss, 1975). Schließlich unterscheiden sich Übereinstimmungsmaße danach, ob in ihnen eine Korrektur für Zufallsübereinstimmung vorgenommen wird oder nicht (s.o.), und Fleiss weist darauf hin, daß die Art dieser Korrektur in π und κ nur eine von mehreren möglichen ist.

Einen etwas anderen Akzent setzen Landis & Koch (1977, 1977a). Sie schätzen nicht nur die Größe von Übereinstimmungen, sondern prüfen an einem oder simultan mehreren Datensätzen mehrere Hypothesen über das Zustandekommen von Daten einer Reliabilitätsstudie mit Hilfe eines allgemeinen Ansatzes zur inferenzstatistischen Analyse von multivariaten kategorialen Daten. Dieses Vorgehen ermöglicht es u.a., gezielt die Ursachen fehlender Übereinstimmung zu identifizieren, wenn z.B. detailliert festgestellt wird, die Beobachter A und B unterschieden sich signifikant im Gebrauch der Kategorien X und Y bei der Teilstichprobe s (vgl. auch Bergen 1980). Abschließend sei auf Trippi & Settle (1976) hin gewiesen, die eine nichtparametrische Variante der internen Konsistenz entwickeln, sowie auf Kaye (1980), dessen Ansatz besonders für die Auswertung *sequentieller* Beobachtungen relevant ist.

Man kann die Frage aufwerfen, welche Relevanz diese Vielzahl von Möglichkeiten für die Forschungspraxis hat, ob also die Unterscheidungen zwar von theoretischem Interesse sind, jedoch - insbesondere bei großen Stichproben - die Schlußfolgerungen gleich sind, die aus verschiedenen Übereinstimmungsmaßen gezogen werden. Zu dieser Frage ist uns keine Literatur bekannt, so daß pragmatische Erwägungen hinter dem Versuch zurückstehen müssen, die theoretisch angemessene Form der Prüfung einzusetzen.

5.2 Übereinstimmungsmaße für ordinalskalierte Daten

Wenn Beobachter ihre Ergebnisse als Rangreihen mitteilen oder Forscher Beobachtungen als Rangreihen auffassen, dann läßt sich nicht prüfen, ob Beobachter in ihren Urteilen über die absolute Merkmalsausprägung übereinstimmen; Unterschiede in Skaleneinheit und Skalensprung können sich bei allen im folgenden besprochenen Übereinstimmungsmaßnahmen nicht mindernd auswirken. Vielmehr sind alle Transformationen der Skaleneinheiten zugelassen, welche die Ordnung der betrachteten Größen nicht verändern. Berechnet man die Übereinstimmung zwischen 2 Rangreihen, so läßt sich Abweichung

entweder beschreiben als Summe der quadrierten Differenzen, wie dies z.B. in Spearmans Rangkorrelationskoeffizient Q geschieht, oder als Summe der absoluten Differenzen der Wertepaare, so in Spearmans (1906) „footrule“ (vertiefte Darstellung in Hubert, 1979). Schon Spearman wies darauf hin, daß eine Quadrierung größere Abweichungen stärker gewichtet als kleinere. Wenn man erwartet, daß extremere Meßwerte mehr Meßfehler enthalten als mittlere, hat der Meßfehler bei Maßen wie Q einen stärkeren Einfluß auf die Bestimmung der Konsensstärke. Spearmans Q bezieht sich auf Rangdifferenzen, was trivialerweise impliziert, jeder Rangplatz sei von seinem jeweiligen Nachbarn gleich weit entfernt. Kruskal (1958) hat jedoch auf eine Interpretationsmöglichkeit von Q hingewiesen, die ohne diese Annahme auskommt: Wenn man Paare von Tripeln vergleicht (je ein Tripel stammt von einem Beobachter und stellt eine Rangordnung von drei Beobachtungen dar), läßt sich Q als Maß für proportionale Fehlerreduktion interpretieren.

Bei der Auswahl zwischen verschiedenen Maßen für ordinale Übereinstimmung kann man die Frage beachten, wie Rangplatzbindungen (ties) behandelt werden (zur gesamten Diskussion s. Hildebrand et al., 1977). In dem von Goodman & Kruskal (1954) vorgeschlagenen γ werden alle Rangplatzbindungen nicht beachtet, während sie in Somers' d_{xy} berücksichtigt werden. Wilson (1974) schließt nur die Beobachtungspaare aus, die auf beiden Variablen rangplatzgebunden sind.

Will man die ordinale Übereinstimmung zwischen mehr als zwei Beobachtern ermitteln, so berechnet man üblicherweise den Konkordanzkoeffizienten W von Kendall (1948). Er wird Null bei maximaler Nichtübereinstimmung und Eins bei völliger Übereinstimmung, d.h. wenn Rangreihen identisch sind. Er läßt sich auf Überzufälligkeit prüfen. Bei m Beurteilern besteht folgende Beziehung:

$$\text{mittleres } Q = \frac{mW - 1}{m - 1} .$$

Ein niedriges W schließt nicht aus, daß die Gesamtmenge der Beobachter in 2 oder mehr Mengen unterteilt werden kann, die jeweils untereinander stark übereinstimmen. Während W die mittlere Übereinstimmung aller Rangreihen untereinander beschreibt, kann Konsens auch geprüft werden als Übereinstimmung mit einem Kriterium, d.h. einer vorgegebenen Rangordnung (Lyerly, 1952; Cureton, 1958, 1965; Taylor & Fong, 1963; Taylor, 1964). Unterschiede in der Konsensstärke zwischen unabhängigen Beobachtergruppen lassen sich inferenzstatistisch prüfen (s. Stewart et al., 1979, S. 310ff.). Schucany & Frawley (1973) legen ein Prüfverfahren für folgende Fragen vor: Besteht innerhalb einer Gruppe von Beobachtern genügend Übereinstimmung? und wenn ja: Besteht zugleich signifikante Übereinstimmung zwischen zwei Gruppen von Beurteilern? (Für Weinkenner sei in diesem Zusammenhang auf Amerine & Roessler, 1976; hingewiesen.)

5.3 Übereinstimmungsmaße für intervallskalierte Daten

5.3.1 Einfache varianzanalytische Ansätze und Intraklassen-Koeffizienten

Wir behandeln zunächst die von Ebel (1951) eingeführte varianzanalytische Auswertung von Reliabilitätsstudien einschließlich der dabei meistens vorgenommenen Bestimmung eines Intraklassen-Korrelationskoeffizienten (Über-sichten z.B. Haggard, 1958; Maxwell & Pilliner, 1968; Landis & Koch, 1975; Bartko, 1976; Kraemer & Korner, 1976; Werner, 1976; Asendorf & Wallbott, 1979). Wenn jeder von d Beobachtern einmal ein Verhaltensmerkmal an allen n Personen beobachtet, läßt sich folgendes *Modell* für das Zustandekommen einer Beobachtung J_{ij} an Person i durch Beobachter j aufstellen (hier nach Landis & Koch, 1975):

$$(1) \quad J_{ij} = \mu + s_i + e_{ij},$$

wobei μ den Gesamtmittelwert, also die mittlere Beobachtungsausprägung, s_i den Effekt des Merkmalsträgers i und e_{ij} den verbleibenden Residualwert oder Fehler wiedergibt. Für ein inferenzstatistisches Vorgehen muß man die Annahmen eines varianzanalytischen Modells mit Zufallseffekten treffen: Die n Personen stellen eine Zufallsstichprobe aus einer angegebenen Population dar; die s_i sind normalverteilt mit einem Mittelwert = 0 und der Varianz σ_s^2 ; die e_{ij} sind normalverteilt mit einem Mittelwert = 0 und der Varianz σ_e^2 ; die s_i und e_{ij} sind unabhängig voneinander. Selva (1976) diskutiert diese Annahmen und einige Auswege, wenn sie nicht erfüllt sind. Der Intraklassen-Koeffizient ist für (1) definiert als

$$(2) \quad Q = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_e^2} \quad ,$$

also als Verhältnis der Varianz zwischen den Merkmalsträgern zur Gesamtvarianz. Um Begriff und Namen des Intraklassen-Koeffizienten zu verstehen, kann man von dem Beispiel ausgehen, in dem der korrelative Zusammenhang zwischen den IQs der Personen eines Zwillingspaares bestimmt werden soll.

Ein Produkt-Moment-Korrelationskoeffizient r beruht auf Paaren von Daten, je ein Datum wird dem Vektor X , das andere dem Vektor Y zugeordnet. Welchen Zwilling ordnet man X , welchen Y zu? r kann je nach Zuordnung erheblich schwanken. Werner (1976, S. 489): „Allgemein tritt dieses Problem immer dann auf, wenn Objekte zwar in Klassen einteilbar sind, aber innerhalb derselben nicht weiter unterschieden werden sollen, und man generell ermitteln will, ob Objekte *einer* Klasse einander ähnlicher sind als Objekte *verschie-dener* Klassen.“

Modell 1 berücksichtigt nicht explizit Unterschiede zwischen den Beobachtern. Diese kann man als zufällige oder als fixierte Beobachtereffekte einführen. Faßt man die Beobachter als eine Zufallsstichprobe aus einer größeren Population potentieller Beobachter auf, so ergibt sich Modell

$$(3) \quad J_{ij} = \mu + s_i + d_j + e_{ij},$$

wobei d_j zusätzlich den Effekt des Beobachters j repräsentiert. Die d_j sind normalverteilt mit Mittelwert = 0 und Varianz σ_d^2 ; die s_i , d_j und e_{ij} sind voneinander unabhängig. Der diesem Modell entsprechende Intraklassen-Koeffizient ist

$$(4) \quad \rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2 + \sigma_e^2} \quad ,$$

wobei σ_d^2 die Varianz darstellt, die auf Unterschiede zwischen den Beobachtern zurückgeht, also „interobserver bias“ darstellt. Faßt man die Beobachter als fixiert auf und interessieren nur Aussagen über ihre Verlässlichkeit, weil gerade sie in weiteren Studien eingesetzt werden sollen, so ergibt sich ein (3) vergleichbares „mixed model“. Wenn die Varianz zwischen den Beobachtern ausgeklammert werden soll, gibt es zwei Möglichkeiten. Bei Ebel, Haggard und Winer (1971) findet sich die entsprechende Varianzkomponente weder im Zähler noch im Nenner der Reliabilitätsformel, bei Rajaratnam (1960) und Krippendorff (1970, s. Werner, 1976) bleibt die Gesamtvarianz im Nenner; für die sich dann ergebenden Schätzprobleme und für die Bestimmung von Konfidenzintervallen s. Lu (1971) sowie Fleiss & Shrout (1978). Werner weist darauf hin, daß der Intraklassenkoeffizient nur dann dem über alle Beobachterpaare gemittelten Produkt-Moment-Korrelationskoeffizienten gleich ist, wenn die Mittelwerte und Varianzen aller Beobachter gleich sind und bei der Berechnung des Intraklassen-Koeffizienten die Varianz zwischen den Beobachtern nicht zur Fehlervarianz geschlagen wird.

Eine ausführliche Ableitung und Begründung der varianzanalytischen Reliabilitätsprüfung für den univariaten Fall ohne Replikation der Beobachtung findet sich in Winer (1971, S. 283ff.). Winer verdeutlicht die beiden Entscheidungen, die der Forscher für seine Reliabilitätsaussage zu fallen hat: Er kann die Verlässlichkeit entweder einer typischen einzelnen oder der über alle Beobachter gemittelten Messung bestimmen wollen, und zwar in beiden Varianten entweder mit oder ohne Einbezug der Varianz zwischen den Beobachtern in die Fehlervarianz. Die Wahl zwischen diesen beiden Alternativen wird er begründen wollen, wozu er die folgenden Überlegungen heranziehen könnte: Die Ausgangsfrage ist stets, zu welchem praktischen oder wissenschaftlichen Zweck er die Beobachtungen verwendet. Hängt beispielsweise die Einführung von Produkten auf einen Markt von den Schätzurteilen mehrerer Experten ab, und werden diese Urteile zusammengefaßt, ist zweifellos die Reproduzierbar-

keit der durchschnittlichen Schätzwerte gefragt. Wenn der Beurteiler jedoch typischerweise einzeln arbeitet, wie z.B. Lehrer bei der Benotung ihrer Schüler, ist die Verlässlichkeit der individuellen Urteile von Interesse. Auch die Entscheidung über die Einbeziehung der Varianz zwischen den Beurteilern in die Fehlervarianz hängt vom Verwendungszweck ab. Immer dann, wenn Unterschiede zwischen den über alle Beobachtungen berechneten Mittelwerten der einzelnen Beurteiler nicht zu Unterscheiden in den Entschlüssen führen, zu denen diese Beobachtungen die Basis liefern, sollte die Varianz zwischen den Beurteilern nicht zur Fehlervarianz geschlagen werden. Dies sollte hingegen wohl geschehen, wenn -wie etwa bei der zentralen Vergabestelle in ihren Entscheidungen über Studienplätze - Noten verschiedener Schüler, die von verschiedenen Lehrern stammen, gegeneinander aufgerechnet oder Durchschnittswerte verglichen werden, die von verschiedenen Beurteilergruppen stammen.

Die detaillierte Information über die Reproduzierbarkeitsstruktur der Daten wird bei dieser Art Analyse erkaufte durch starke Modellannahmen: (1) der Meßfehler korreliert nicht mit dem wahren Wert; (2) die Stichprobe der beobachteten V_{pn} ist eine Zufallsstichprobe aus der Population der Personen, auf die Schlüsse gezogen werden sollen; (3) die Stichprobe der Meßinstrumente oder Beobachter, welche Replikationen bereitstellen, ist in Modell (3) eine Zufallsstichprobe vergleichbarer Instrumente oder Beobachter; (4) die Schätzung der Fehlervarianz basiert auf der Zusammenfassung der Varianzen innerhalb jeder V_p . Die Datenerhebungssituation muß auf die Bedingungen (2) und (3) zugeschnitten werden, andere Modellimplikationen lassen sich nach der Erhebung prüfen: Da der „wahre Wert“ über alle Beobachter konstant bleibt, muß die Korrelation zwischen den verschiedenen Beobachtern statistisch konstant sein; die entsprechende Varianz-Kovarianzmatrix ließe sich auf Homogenität prüfen.

Wenn jeder Beobachter r wiederholte Einschätzungen des gleichen Verhaltens an den gleichen Personen vornimmt, läßt sich (3) für die k -te Replikation erweitern auf

$$(5) \quad J_{ijk} = \mu + s_i + d_j + (sd)_{ij} + e_{ijk},$$

wobei $(sd)_{ij}$ die Interaktion zwischen Beobachter j und Merkmalsträger i darstellt und ein Maß dafür ist, wie sehr j bei der Beurteilung von i von seinem üblichen Reaktionsmuster abweicht. Will man nicht, wie alle bisher vorgeschlagenen Modelle, annehmen, die Varianzkomponente zu Lasten des Zufallsfehlers sei für alle Beobachter gleich, so kann man auf ein Modell von Grubbs (1948, 1973) zurückgreifen (s.a. Overall, 1968 für die Situation, in der n Personen auf m Bedingungen aufgeteilt und anschließend von zwei unabhängigen Beobachtern beurteilt werden). Beurteilt jeder Beobachter mehr als ein Merkmal, so kommt eine multivariate Erweiterung des varianzanalytischen

Ansatzes von Fleiss (1966) in Frage, allerdings insbesondere auch die später besprochenen Generalisierbarkeitsstudien (zur Diskussion, wie die Residualvarianz zu schätzen sei, s. Huck 1978).

Diesen Abschnitt abschließend sollen noch einige Varianten besprochen werden. Lawlis & Lu (1972) schlagen einen bei Tinsley & Weiss (1975) kritisch besprochenen Index vor, der es erlaubt, Übereinstimmung als fehlerfreien Fall festzulegen, aber auch als Abweichung bis zu einer, bis zu zwei, etc. Maßeinheiten der gewählten Skala. Finn (1970, 1972) weist darauf hin, daß bei varianzanalytischen Reliabilitätsschätzungen Beobachtungen über mehr als eine Person oder Verhaltensweise vorliegen müssen, und die Varianz zwischen den Personen oder Verhaltensweisen muß ausgeprägt sein, damit sich ein genügend großer Koeffizient ergeben kann. Situationen sind jedoch nicht selten, in denen nur eine Gegebenheit, diese jedoch mehrfach beobachtet wird, und für diesen Fall ist Finns Methode gedacht (zur Kritik insbesondere an Finns Inferenzstatistik s. Tinsley & Weiss, 1975).

Dem varianzanalytischen Modell folgend hat Krippendorff (1970) einen Ansatz vorgelegt, der insbesondere die während der Konstruktionsphase von Kategoriensystemen wichtige Frage nach Art und Größe der einzelnen Quellen für fehlende Verlässlichkeit zu beantworten gestattet. Globale Gesamtwerte geben keine Information über mögliche Schritte zur Verbesserung. - Detaillierte Informationen können erwünscht sein über

- (1) die geschätzte Reliabilität einer Datenmenge über alle Datenquellen. Man könnte diese Schätzung *Datenreliabilität* nennen und als Maß für das generelle Vertrauen in der Daten interpretieren;
- (2) das geschätzte Ausmaß, in dem sich die Datenreliabilität verbessern ließe, wenn die Urteile einiger Beurteiler transformiert oder neu definiert würden. Diese Schätzung erfaßt den *systematischen Fehler* des Verfahrens und ergibt zusammen mit dem **Zufallsfehler** den Betrag, der bis zu einer völligen Datenreliabilität fehlt;
- (3) für jeden einzelnen Beobachter das geschätzte Ausmaß seiner Verlässlichkeit, also die *individuelle Reliabilität*. Mit dieser Information ließe sich feststellen, ob - und, wenn ja, welche - einzelne Beobachter besonders unzuverlässige, d.h. hier: vom allgemeinen Trend der Beobachtergruppe abweichende Werte liefern. Je nach Fragestellung könnte dann die Schulung dieser Beobachter verbessert oder die Daten dieser Beobachter könnten nicht ausgewertet werden;
- (4) für jeden einzelnen Beobachter eine Schätzung des Ausmaßes, in dem seine Beobachtungen durch Schulung oder Transformation korrigierbar sein würden. Geschätzt werden müßte also der *systematische individuelle Beobachterfehler*, der zusammen mit dem *individuellen Zufallsfehler* die individuelle Unreliabilität ausmacht;
- (5) für jede einzelne Kategorie (recording unit) eine Schätzung des Ausmaßes,

in dem eine Zufallsstichprobe von Beobachtern bei ihren Beobachtungen zu dieser Einheit übereinstimmt. Diese Schätzung der *Reliabilität einer Kategorie* ermöglicht die gezielte Modifizierung oder Eliminierung einzelner Kategorien eines Systems.

Das Verfahren von Krippendorff erlaubt, diese Fragen zu beantworten. Sommerbe (o.J.) hat auf Druckfehler in der Arbeit von Krippendorff hingewiesen und ein Computerprogramm bereitgestellt.

5.3.2 Generalisierbarkeitsstudien

Generalisierbarkeitsstudien (Cronbach et al. 1963, 1972; Gleser et al. 1965) verallgemeinern den varianzanalytischen Ansatz; man kann diese Studien als Anwendung der Prinzipien multifaktorieller und multivariater Varianzanalysen auf die Reliabilitätsproblematik auffassen. Alle Generalisierbarkeitskoeffizienten sind Intraklassen-Koeffizienten; sie sind Maße der Verallgemeinerbarkeit einer Beobachtung über verschiedene Entstehungsbedingungen. Die Generalisierbarkeitstheorie stellt insofern gegenüber der Reliabilitätskonzeption in der klassischen Testtheorie eine „Liberalisierung“ dar (s. Kristof in diesem Band), als nicht gefordert wird, die Beobachtungen unter verschiedenen Bedingungen müßten „parallel“ sein, d.h. gleiche Mittelwerte, Varianzen und Kovarianzen aufweisen.

Für die Untersuchung der Reproduzierbarkeit von Beobachtungen kann man folgende Punkte als wesentliche Fortschritte gegenüber dem einfachen varianzanalytischen Ansatz herausstellen: (1) die ausdrückliche Festlegung der Facetten, auf die verallgemeinert werden soll, (2) die Hinweise auf den Zusammenhang zwischen Aussageabsicht und Definition des Meßfehlers, und (3) die ausdrückliche Unterscheidung zwischen Generalisierbarkeits- oder G-Studien und Entscheidungs- oder D-Studien, womit der Zweck, den die Reliabilitätsanalyse haben soll, thematisiert wird, nämlich eine wissenschaftliche oder praxisbezogene Entscheidung aufgrund der Beobachtungen zu fällen.

Cronbach et al. (1972, S. 15) formulieren den Grundsatz folgendermaßen: „Der Meßwert, auf den die Entscheidung sich stützen soll, ist nur einer von vielen Meßwerten, die den gleichen Zweck erfüllen könnten. Derjenige, der die Entscheidung zu treffen hat, interessiert sich fast nie für die Reaktion, die auf ein ganz bestimmtes Reizobjekt oder auf die speziellen Fragen, gegenüber einem ganz bestimmten Versuchsleiter oder in dem einen besonderen Augenblick der Testdurchführung gegeben werden. Wenigstens einige dieser Meßbedingungen könnten variiert werden ohne den Meßwert weniger akzeptabel für denjenigen zu machen, der die Entscheidung zu fällen hat. D.h., es gibt ein Universum von Beobachtungen, und jede dieser Beobachtungen hätte eine brauchbare Basis für die Entscheidung abgegeben. Der ideale Wert, auf den

man die Entscheidung basieren könnte, wäre der Meßwert einer Person, gemittelt für alle akzeptierbaren Beobachtungen, den wir ihren ‚Universalwert‘ (Universe score) nennen wollen. Der Untersucher benutzt den beobachteten Wert oder eine Funktion des beobachteten Wertes als ob das der Universalwert wäre. Das heißt, er generalisiert von der Stichprobe auf das Universum.“ *Die Frage nach der ‚Reliabilität‘ entwickelt sich somit zu einer Frage nach der Genauigkeit der Generalisation oder der Verallgemeinerbarkeit.* Wenn man die Verlässlichkeit von Beobachtungen prüfen will, wäre es günstig, möglichst im vollständig überkreuzten Versuchsplan die Bedingungen Beobachter, Kategorien, Personen und Zeitpunkte einzuführen. Dann läßt sich auch die von Kraemer & Korner (1976) beklagte Vermischung von interindividuellen Unterschieden, Stabilität des Merkmals, intraindividuelle Konsistenz jedes Beobachters und interindividuellen Unterschieden zwischen verschiedenen Beobachtern entmischen. Ein oft erwünschtes und einfaches Ergebnis läge dann vor, wenn der Haupteffekt zu Lasten der Beobachter und alle Interaktionen, an deren Definition die Beobachter beteiligt sind, keine nennenswerte Varianz auf sich vereinigen. Dann könnte man annehmen, es bestünden keine Unterschiede zwischen den Beobachtern in ihren allgemeinen Reaktionstendenzen oder in ihren Skalenverankerungen, sie benützten die Kategorien gleichartig, beurteilten die Personen nicht idiosynkratisch und schwankten in ihrem Beobachtungsverhalten nicht über verschiedene Zeitpunkte (zur Verbindung von Kosten-Nutzen-Analysen bei der Planung von Beobachtungsstudien mit Generalisierbarkeitsanalysen s. Gleser et al. 1965; für die Weiterentwicklung der multivariaten G-Studien z.B. auf Profilreliabilität oder Verallgemeinerbarkeit von gain scores s. Joe & Woodward, 1976; Anwendungsbeispiele: Rudinger & Feger, 1970; Levy, 1974; Feger, 1978; Weiterentwicklung bei Nußbaum, 1980).

5.3.3 Pfadanalytische Modelle für die Reliabilitätsprüfung

Noch „liberaler“, d.h. allgemeiner gegenüber dem einfachen varianzanalytischen Ansatz als die Generalisierbarkeitstheorie ist die pfadanalytische Prüfung von Modellen der Datenstruktur, die Annahmen über den Meßfehler einschließt. Der Ansatz von Jöreskog erfordert nicht einmal die Annahme, verschiedene Beurteiler würden gleiche Meßeinheiten benutzen (Werts et al., 1974). Wesentlicher jedoch als die Flexibilität scheint mir die Möglichkeit zu sein, Voraussetzungen prüfen zu können, die den verschiedenen Ansätzen, auch dem Intraklassen-Koeffizient, zugrunde liegen. So läßt sich die Annahme prüfen, allen Messungen liege der gleiche wahre Wert τ zugrunde. Läßt sich diese Annahme nicht halten, dann ist die Bedeutung einer varianzanalytischen Reliabilitätsschätzung unklar. Möglicherweise liegt nicht nur ein „wahrer“ Faktor zugrunde, und die Annahmen über die Unabhängigkeit der Fehler können falsch sein. Dann läßt sich weiter prüfen, ob die Annahme sowohl der

Intraklassen-Korrelation als auch der Generalisierbarkeitstheorie gilt, die Maße verfügten alle über die gleiche Maßeinheit, d.h. „essentially tau equivalent“ (Lord & Novick, 1968, S. 50) sind. „If this hypothesis is rejected then the Anova formulation is rejected whether used for estimating reliability or for generalizability procedures.“ (Werts et al., 1974, S. 29). Weiter läßt sich die im Intraklassen-Koeffizient implizierte Annahme prüfen, die Meßfehler der einzelnen Messungen seien unabhängig.

Für die pfadanalytische Behandlung der Reliabilitätsproblematik sind folgende Schritte charakteristisch: 1. Man nimmt an, jede Messung sei aus einem wahren Wert und einem Meßfehler zusammengesetzt. 2. Man legt die Art des Zusammenhanges zwischen wahren Wert und Meßfehler jeder Messung mit wahren Wert und Meßfehler jeder anderen Messung in einem Modell für die gesamte Datenmenge explizit fest; das pfadanalytische Vorgehen ist in dieser Hinsicht, wie die Art des Zusammenhanges festgelegt wird, sehr flexibel. 3. Um die Parameter in diesem Modell bestimmen und das Modell auf seine Anpassungsgüte prüfen zu können, genügt nicht eine Messung nur einer Variablen. Auf irgendeine der folgenden Weisen muß repliziert werden. Verschiedene Arten der Replikation über Person oder Beobachtungsgegebenheiten lassen sich auch miteinander verbinden. Wir unterscheiden folgende Möglichkeiten: (A) Eine Variable wird zu zwei oder mehr Zeitpunkten erhoben (Retest-„stability“). (B) Eine Variable wird zum gleichen Zeitpunkt durch verschiedene Operationalisierungen erfaßt (interne Konsistenz, hier oft als „reliability“ bezeichnet). (C) Zwei oder mehr inhaltlich verschiedene Variablen werden an den gleichen Personen erhoben, meistens in Verbindung mit A oder B. Werden A und B verbunden, kann man die Beziehungen zwischen Messungen und Konstrukten differenziert spezifizieren - „measurement specification“; wird C mit A oder B oder beiden kombiniert, so lassen sich die Beziehungen zwischen den Konstrukten spezifizieren - „theoretical specification“. Literatur zu A: Heise, 1969; Wiley & Wiley, 1971; Werts et al., 1971; Wheaton et al., 1977; Anwendungsbeispiel: Feger, 1978. Literatur zu A und B: Costner, 1969, Blalock, 1970; Hauser & Goldberger, 1971. Zu C mit A, teils auch mit B: Bohrnstedt, 1969; Duncan, 1969; Heise, 1970. Probleme der theoretischen und der Meßspezifikation gleichzeitig behandeln: Duncan, 1972, 1975, Hannan et al., 1974; Jöreskog & Sörbom, 1976; Mayer & Younger, 1974; Wheaton et al., 1977.

Bei der pfadanalytischen Untersuchung von Meßfehlern wird oft ausdrücklich unterschieden zwischen dem zu erfassenden Konstrukt, bisweilen als unobserved variable bezeichnet, und den - meßfehlerhafteten - Beobachtungen. Jacobson & Lulu (1974) unterscheiden drei Methoden, wie man die durch Meßfehler unverfälschten Beziehungen zwischen Konstrukten bestimmen kann, über einen *einfachen Indikator* (Single indicator), über einen *Index* und über *mehrfache Indikatoren* (multiple indicators). Arbeitet der Forscher nur

mit einem Indikator, so muß er annehmen, dieser erfasse das Konstrukt in dem Sinne gut, als er den größten Teil der Variation der wahren Werte einfange, und ferner bestehe kein Spezifikationsfehler im Modell: „In other words, given two theoretical variables, both of which are measured by single indicators, these two indicators are assumed to be associated *only* through the posited relationship linking the two theoretical variables. If the latter condition cannot be reasonably assumed - and in most practical situations it cannot - estimates of the structural parameters will be biased even when the first assumption holds.“ Ein Index wird aus mehreren einzelnen Indikatoren als Gesamtwert gebildet. Dabei wäre theoretisch begründet zu entscheiden, wieviele und welche Indikatoren kombiniert werden, welches Gewicht jeder einzelne Indikator erhält, und welcher Art, z.B. additiv, die Zusammenfassung sein soll. Arbeitet man mit multiplen Indikatoren, so legt man in einer Hilfstheorie (auxiliary theory) fest, wie die Indikatoren mit bestimmten Konstrukturen verbunden sind; diese Festlegungen sind meistens prüfbar und die Qualität eines jeden Indikators läßt sich bestimmen (s. Costner, 1969; Blalock, 1969; Mayer & Younger, 1974). Beim Vergleich der drei Methoden bevorzugen Jacobson & Lulu aus mehreren Gründen den multiple indicator approach.

5.4 Besondere Erhebungspläne

Keineswegs nur bei intervallskalierten Daten können durch die Art der Erhebung Beobachtungen in einer Ordnung anfallen, die für die Reliabilitätsprüfung besondere Probleme oder Möglichkeiten ergibt. Wir skizzieren zwei Fälle; im ersten liegen die Beobachtungen als Paarvergleichsmatrix über alle Paare von Beobachtungsgegebenheiten vor, im zweiten Fall ist bei der Analyse der Beobachtungen deren Sequenz zu berücksichtigen. Hubert (1979a) hat vier Fragestellungen formuliert, die sich ergeben, wenn *Paarvergleichsmatrizen*, in deren Zellen Ränge stehen, auf Übereinstimmung (concordance) betrachtet werden: 1. Die Übereinstimmung zwischen zwei oder mehr Matrizen - hier: zwischen Beobachtern, die ihre Beobachtungen als Ränge angeben. 2. Der Vergleich mehrerer Matrizen mit einer spezifischen Matrix - hier: z.B. Vergleich aller Beobachter mit einem Standard. 3. Eine Technik zum Vergleich der Übereinstimmung innerhalb und zwischen Teilmengen von Matrizen, z.B. Vergleich von Beobachterteilgruppen untereinander und miteinander. 4. Suche nach einer Matrix, die alle Urteile aller Beobachter am besten repräsentiert. Die Hypothese, die Übereinstimmung sei nur zufällig, wird geprüft, indem man das beobachtete Übereinstimmungsmaß mit der Verteilung derjenigen Übereinstimmungsmaße vergleicht, die bei allen logisch möglichen Permutationen der intakten Zeilen und Spalten einer der verglichenen Matrizen entstehen. Auf diese Weise werden Abhängigkeiten, die zwischen den Werten in den Zellen der Matrizen durch die Rangordnung über alle Zellenwerte einer Matrix entstehen, für die Inferenzstatistik berücksichtigt. Ein Spezialfall ist gegeben,

wenn in den Zellen nicht Ränge, sondern dichotome Werte stehen, wie z.B. bei soziometrischen Wahldaten (dazu: Hubert & Baker, 1978a).

Im zweiten Fall soll bei der Bestimmung der Reliabilität die in den Aufzeichnungen jedes einzelnen Beobachters enthaltene Information über die *Ordnung oder zeitliche Abfolge* der verschiedenen Verhaltensweisen berücksichtigt werden. Dies ist immer dann erforderlich, wenn aus der Reihenfolgeinformation inhaltliche Schlüsse gezogen werden sollen, z.B. in Kausal- und Bedingungsanalysen. Einige Probleme lassen sich an folgender Tab. 3 verdeutlichen (nach Hollenbeck 1978, seine Tab. 6) :

Tabelle 3 : Übereinstimmung in segmentierten Protokollen

verschiedene Zeiteinheiten (sec)	<i>Beobachter 1</i>		<i>Beobachter 2</i>		<i>Beobachter 3</i>	
	Zeit	Ereignis	Zeit	Ereignis	Zeit	Ereignis
1	A		A	A	kein Wert	
2	A	A	A		A	A
3	A		B		A	
4	B		B		B	
5	B	B	B	B	B	B
6	B		B		B	
7	A	A	A	A	C	
8	A		C		C	C
9	C	C	C	C	C	
10	C		C		C	

Das erste Problem, *Synchronisation von Protokollen* (protocol alignment), zeigt sich beim Vergleich des dritten mit den übrigen Beobachtern; er hat den Start verpaßt. Wie läßt sich sein Protokoll mit den übrigen synchronisieren? Und es können auch Ausfälle während der laufenden Beobachtung eintreten. Das zweite Problem besteht in der *Bestimmung der Fehlerart*: Beobachter 3 registrierte in Sek. 7 Verhaltensweise A nicht. Wenn dies sein Fehler war: Hat er das kurze Auftreten übersehen (error of omission, Fehler des Übersehens) oder hat er A mit C verwechselt (error of commission, Verwechslungs-Fehler)? Ein drittes Problem ergibt sich aus dem Vergleich der Zeit- und der Ereignisnotierung für die ersten beiden Beobachter: Die Ereignissequenzen beider Protokolle stimmen völlig überein, die Verhaltensweisen in sec. 3 und 8 nicht. Welche Art Fehler sind diese; und wie soll man sie für die Reliabilitätsbestimmung berücksichtigen? Hollenbeck diskutiert Lösungen für jedes Problem, von denen auch nach seiner Meinung keine zufriedenstellend ist. Der For-

scher, der diesen Erhebungsplan benutzt, sollte zumindest berichten, ob diese Schwierigkeiten in seinen Daten aufgetreten sind und wie er sie gelöst hat (s. auch Kaye 1980).

5.5 Die Berücksichtigung von Reliabilitätskenntnissen bei der weiteren Datenauswertung

Es dürfte selten sein, daß Beobachtungen nicht weiter ausgewertet werden, nachdem sich herausstellt, daß sie ein (vorher festgesetztes) Minimum an Reliabilität nicht aufweisen. Man wird auch nicht allgemein, sondern höchstens im einzelnen Fall festlegen können, wie hoch das erforderliche Minimum anzusetzen wäre. Eine realistische Strategie im Umgang mit Reliabilitätsinformation zeigt sich in den Veröffentlichungen der letzten Jahrzehnte: Die Frage lautet, wie durch Meßfehler und Unzuverlässigkeit der Daten der Schluß - z.B. über die Richtung einer Mittelwertsdifferenz, ihre Signifikanz oder die Höhe einer Korrelation - verfälscht werden könnte, den man aus den Beobachtungen ziehen möchte. Im folgenden können wir nur einige Beispiele erwähnen, die untereinander zwar verschieden sind, aber die Thematik nicht vollständig abdecken.

Unreliable Kategorien und Beobachter führen zu falschen Klassifikationen. Keys & Kihlberg (1963, hier nach Fleiss, 1973 S. 135ff.) behandeln die Frage, wie groß der Fehler ist und in welche Richtung er die Analyse verfälscht: Gegeben seien ein Merkmal, das relativ eindeutig feststellbar ist, z.B. Lungenkrebs, und ein anderes, das weniger sicher zu klassifizieren ist, beispielsweise Häufigkeit des Rauchens. P_L sei die Proportion von Lungenkrebspatienten, die häufig rauchten, somit ist $1-P_L$ die Proportion der Patienten, die nicht rauchten. E_L sei die Proportion der Patienten, die zwar rauchten, aber fälschlich als Nichtraucher klassifiziert wurden, und F_L die Proportion der Patienten, die fälschlich als Raucher eingestuft wurden. Statt wie gewünscht P_L bestimmen zu können, ist nur p_L , die beobachtete Proportion der rauchenden Patienten, aus den Angaben der Personen oder Beobachter zu erhalten. Dabei ist

$$p_L = (1-E_L)P_L + F_L(1-P_L).$$

Also setzt sich p_L additiv aus zwei Teilen zusammen, aus wahren Rauchern und wahren Nichtrauchern. Ob $p_L = ,>, <P_L$ hängt von den beiden Fehlerraten F_L und E_L ab:

$$p_L > P_L \text{ wenn } \frac{F_L}{E_L + F_L} > P_L \quad ,$$

$$p_L = P_L \text{ wenn } \frac{F_L}{E_L + F_L} = P_L \quad ,$$

$$p_L < P_L \text{ wenn } \frac{F_L}{E_L + F_L} < P_L \quad .$$

Nicht nur die Schätzung der wahren Proportion wird verständlicherweise durch Unreliabilität beeinträchtigt, sondern auch der *Vergleich zwischen zwei* Proportionen: Ein Forscher möchte prüfen, ob die Proportion der rauchenden Lungenkrebspatienten $P_B < P_L$ ist, und bestimmt die Differenz D zwischen den wahren Proportionen:

$$D = P_L - P_B$$

Beobachtet ist jedoch

$$d = D + (F_L - F_B) + P_B(E_B + F_B) - P_L(E_L + F_L),$$

und d kann größer, gleich oder kleiner D sein. Wenn

$$E_L = E_B = E \text{ und } F_L = F_B = F,$$

ist $d = D(1 - (E + F))$.

Hieraus ergibt sich: (1) Die beobachtete Differenz kann nicht gleich der wahren Differenz sein, wenn auch nur eine Fehlerrate ungleich Null ist. (2) Wenn beide Fehlerraten unter 50% liegen, hat d die gleiche Richtung wie D , ist jedoch numerisch kleiner. Dies ist der seit Bross (1954) bekannte Minderungseffekt, der jedoch - wie die eben skizzierte Analyse ergab - keineswegs immer auftritt. (3) Die Annahme ist falsch, Unreliabilität könne lediglich Unterschiede mindern; sie kann auch die Richtung eines Zusammenhanges fälschen (weitere Literatur zur Auswirkung von Fehlklassifikationen auf Zusammenhangsmaße und χ^2 : Rogot, 1961; Mote & Anderson, 1965, Assakul & Procter, 1967; Koch, 1969). Fleiss (1973) behandelte ausführlich und mit zahlreichen Literaturverweisen die Frage, wie der Meßfehler statistisch kontrolliert und wie Korrekturfaktoren bestimmt werden können.

Ein weiterer Forschungsschwerpunkt in diesem Bereich befaßt sich mit der Frage, wie Unreliabilität den Typ II oder β -Fehler oder die Prüfstärke ($1 - \beta$, power, s. Cohen, 1969; Bredenkamp, 1972) eines Inferenztests beeinflusst. Nicewander & Price (1978) haben kürzlich den Stand der Diskussion zusammengefaßt. Die Annahmen über den Meßfehler sind dabei in der Regel die der klassischen Testtheorie, die über den Stichprobenfehler jene des allgemeinen linearen Modells, speziell der Varianzanalyse. Der Fehlerterm der Varianzana-

lyse besteht dann aus der Abweichung des wahren Wertes vom Bedingungs-mittelwert plus dem Meßfehler. Ändert man die Reliabilität der abhängigen Variablen, so kommt es für die Auswirkung auf die Prüfstärke darauf an, ob dadurch die Varianz der wahren Werte, der Fehler, oder beide geändert wurden. Nimmt man wie Sutcliffe (1958) und Cleary & Linn (1969) an, nur die Varianz der Fehler werde verringert, dann muß die Prüfstärke ansteigen. Sutcliffe (1980) zeigt erneut den nachteiligen Effekt der Unreliabilität von Differenzen und der abhängigen Variablen für die power einer statistischen Prüfung.

Abschließend seien noch einige weitere Beispiele aufgezählt: Werts & Linn (1971) zeigen, wie man die Unreliabilität einer Kovariaten in einer Kovarianzanalyse berücksichtigen kann. Rock et al. (1977) behandeln den Fall des linearen Modells, wenn (auch) die unabhängigen Variablen fehlerbehaftet sind. Stroud (1974) diskutiert Möglichkeiten des Posttest-Vergleichs bei zwei Gruppen, wenn fehlerbehaftete Pretest-Werte vorhanden sind, und Alternativen zur üblichen Kovarianzanalyse. Nur bei Unreliabilität können Regressionsartefakte auftauchen, die Hunter & Cohen (1974) bei der Analyse nichtlinearer Modelle der Einstellungsänderung berücksichtigen. Huber (1973) hat die Bedeutung der testspezifischen Fehlervarianz für die zufallskritische Einzelfalldiagnostik herausgearbeitet. In der Entscheidungsforschung hat Fischer (1976) zwei verschiedene Fehlertheorien, eine für rating, die andere für Paarvergleiche inhaltlich begründet und in ihren Auswirkungen auf die Prüfung des multidimensionalen Nutzenmodells analysiert. Cochran (1968) gibt eine gute Einführung in den gesamten Bereich.

6. *Validität von Beobachtungen*

Man kann zur Beobachtung Apparate verwenden und ihre Daten statistisch analysieren, ohne daß ein Mensch dazwischentritt; ein Beobachter hat also keinen Einfluß auf die Daten. Dann könnten die Beobachtungen völlig objektiv und reliabel sein; das Validitätsproblem stellt sich dennoch, denn aus den Beobachtungen werden Schlüsse gezogen. Diese Schlüsse können sich von einer Menge von Beobachtungen auf eine andere Menge beziehen oder auf theoretische Vorstellungen, und als Vorhersagen oder Korrespondenzaussagen formuliert sein. Solche Schlüsse können für jeweils bestimmte Beobachtungen und bestimmte theoretische Vorstellungen gültig sein. Validität - oder synonym Gültigkeit - bezieht sich zunächst auf Schlußfolgerungen: Beobachtungen oder bestimmte Verfahren, Beobachtungen zu gewinnen, kann man dann als valide bezeichnen, wenn die auf ihnen basierenden Schlüsse zutreffen. Die Methodenlehre der Validierung, d.h. der Prüfung von Validität, befaßt sich mit der Frage nach den Kriterien (und ihrer Anwendung) der Prüfung, ob diese Schlüsse zutreffen. Hierzu wurden verschiedene Kriterien und Verfahrensweisen entwickelt, die in den folgenden Abschnitten geschildert werden.

Wenn Beobachtungen von menschlichen Beobachtern stammen, kommen nicht nur Täuschung, Selbsttäuschung und systematisch verfälschende Beobachtungstendenzen als problematisch für die Gültigkeit hinzu. Vielmehr wird man menschlichen Beobachtungen, läßt man die Möglichkeit von Täuschungen und Fälschungen einmal beiseite, im Prinzip auch Gültigkeit zusprechen müssen, ohne daß sich die Gültigkeit auf einen logischen Schluß bezieht. Der Autofahrer - um bei Alltagsbeobachtungen zu bleiben - sieht, daß die Ampel rot ist, der Vordermann doch noch durchfährt, schlägt mit der Hand an seine Stirn und ärgert sich. Die Beobachtungen und Schilderungen des eigenen und fremden Verhaltens und des eigenen Erlebens kann man als valide bezeichnen in dem Sinn, daß sie sich auf die unmittelbare Anschauung des Beobachters beziehen. Diese Basis, die jede menschliche Beobachtung konstituiert, wird im folgenden nicht weiter besprochen. Zwar kann eine Methodenlehre Bedingungen und Prüfmöglichkeiten angeben, wann und wie Fremd- und Selbstbeschreibungen verfälscht werden; sie muß die grundsätzliche Möglichkeit gültiger Beobachtungen durch Menschen jedoch voraussetzen. Validitätstheorie befaßt sich also nicht mit den philosophischen Bedingungen der Möglichkeit von - zutreffenden - Beobachtungen durch Menschen und der näheren Bestimmung des Wortes „zutreffend“, und auch nicht mit der Fülle physiologischer, psychologischer, sozialer und anderer Faktoren, die verzerrend und störend auf die Prozesse der Wahrnehmung und des Beurteilens einwirken können. Validitätstheorie innerhalb der Methodenlehre befaßt sich mit der Prüfung der Gültigkeit von Schlüssen aus Beobachtungen.

Die Prüfung, ob Schlüsse von Beobachtungen auf Konstrukte und auf andere Beobachtungen zutreffen, kann nach verschiedenen Kriterien geschehen. Diese Kriterien führen zur Unterscheidung verschiedener *Validitätsarten*. Üblicherweise werden folgende Validitätsarten unterschieden: Kriteriumsvalidität mit den Varianten prädiktive und gleichzeitige (concurrent) Kriteriumsvalidität, Inhalts- oder Kontentvalidität und Konstruktvalidität (vgl. Davis, 1974). Bei *Kriteriumsvalidität* interessiert der Schluß vom gerade beobachteten Verhalten auf anderes Verhalten, das gleichzeitig oder zeitlich später gezeigt wird, sei es nun gleichartiges oder verschiedenes Verhalten. Zweck ist die Vorhersage (im statistischen Sinn) einer Menge von Beobachtungen aufgrund einer anderen Beobachtungsmenge.

Die *Inhaltsvalidität* bezieht sich auf einen Repräsentationsschluß, der von einer Stichprobe von Beobachtungen auf ein definiertes Universum gerichtet ist (zur Vertiefung und für weitere Literatur s. Klauer, 1978). Wegen ihrer besonderen theoretischen Wichtigkeit widmen wir der Konstruktvalidität eigene Abschnitte.

6.1 Konstruktvalidierung

Für zahlreiche Variablen, die durch Beobachtung erfaßt werden sollen, läßt sich inhaltliche oder kriterienbezogene Validität jedoch (noch) nicht nachweisen, weil man nicht über ein plausibles Kriterium verfügt oder man das Universum der Verhaltensweisen, Situationen, Aufgaben o.ä. nicht erschöpfend und nach theoretisch fundierten Taxonomien beschreiben kann. Für solche Fälle wird der langwierige Prozeß der **Konstruktvalidierung** erforderlich, dessen Grundgedanken erstmals umfassend von Cronbach & Meehl (1955) systematisiert wurden: „Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim.“ Zwischen verschiedenen theoretischen Konstrukten und zwischen Konstrukten und Beobachtungen bestehen Beziehungen, die in der Konstruktvalidierung expliziert und - soweit möglich - empirisch geprüft werden.

Eine Beziehung stellt, formal betrachtet, den Zusammenhang von mindestens zwei Variablen dar. Der Zusammenhang kann statistisch auf verschiedene Weise analysiert werden, z.B. als Vergleich von Mittelwertdifferenzen bei der einen Variablen, wenn die andere die Gruppierung bedingt, als Trendhypothese, etc. Besondere Bedeutung hat in der Validierungsforschung die Formulierung des Zusammenhanges als Korrelation oder Kovariation gefunden. Besteht zwischen zwei reliablen Meßwertreihen kein Zusammenhang, so ergibt sich kein Problem für die Schlußfolgerung, auf die sich ein Validitätsurteil stützen könnte. Besteht jedoch ein signifikanter und substantieller Zusammenhang, so sind mehrere Schlüsse möglich. 1. Es besteht ein inhaltlich begründeter Zusammenhang zwischen A und B; 2. die experimentelle Manipulation, die B variieren sollte, variierte statt dessen A; 3. zwar wurde B experimentell variiert oder zeigte ohne Eingriff des Forschers Variation, das Instrument jedoch, das A erfassen sollte, registriert in Wirklichkeit B (Tesser & Krauss, 1976). Wenn A und B beispielsweise Ängstlichkeit und Selbstwertgefühl darstellen, könnten alle drei Deutungen a priori gleich plausibel sein. Den Fall, daß eine dritte Variable A und B beeinflusst, verfolgen Tesser & Krauss zunächst nicht weiter. Ihre entscheidende Frage ist dann, unter welchen Bedingungen man schließen kann, der beobachtete Zusammenhang zwischen zwei Variablen bedeute etwas anderes als daß beide Indikatoren des gleichen Konstrukts sind. Dafür stellen sie eine nicht erschöpfende Liste von Kriterien auf. Ein wichtiges Kriterium besteht darin zu beobachten, ob die Veränderung einer dritten Größe die Beziehung zwischen A und B ändert; wenn ja, können sie nicht (nur) Indikatoren des gleichen Konstruktes sein. Ein weiteres Kriterium prüft, ob die Effekte von A auf B und von B auf A asymmetrisch sind, ein drittes Kriterium die Konstanz der Beziehung zwischen A und B über verschiedene Erfassungsmethoden. Die Frage, wie ein Zusammenhang zwischen

zwei Variablen, die unterschiedliche Konstrukte erfassen sollen, zu deuten ist, sollte logisch vor einer Prüfung der konvergenten und diskriminanten Validität geklärt werden.

Die Möglichkeiten der Konstruktvalidierung, sofern diese korrelationsanalytisch vorgeht, wurden durch den Ansatz der *konvergierenden und diskriminierenden Validierung* mit Hilfe einer multitrait-multimethod matrix (MTMM) von Campbell & Fiske (1959) wesentlich vorangetrieben und systematisiert. Der Ansatz geht von folgenden Erwägungen aus: (1) Wenn unabhängige, in ihrer Operationalisierung des Konstruktes sich nicht überlappende Meßverfahren das Gleiche erfassen sollen und gültig sind, so müssen sie, auf die gleichen Meßgegebenheiten angewandt, zu gleichen Ergebnissen kommen; die Ergebnisse müssen konvergieren. (2) Beobachtungen und Messungen, die inhaltlich verschiedene Konstrukte erfassen sollen, dürfen nicht höher korrelieren als aufgrund der Theorie des jeweiligen Gegenstandsbereiches zu erwarten ist, auch dann nicht, wenn sie mit Verfahren gewonnen werden, die technisch identische Operationalisierungsschritte aufweisen. Auch von der Erhebungsmethodik her gesehen gleiche Verfahren müssen inhaltlich verschiedene Konstrukte diskriminieren, wenn die Beobachtungen valide sein sollen. (3) Bei jeder konkreten Beobachtung oder Messung wird eine „trait-method unit“, eine Einheit von spezifischem Inhalt und bestimmtem, allerdings nicht an diesen Inhalt gebundenem Meßvorgehen erfaßt. So liegt nun der Gedanke nahe, die systematische Varianz zwischen Beobachtungen aufzuspalten in einen Teil zu Lasten der Reaktionen auf unterschiedliche Eigentümlichkeiten verschiedener Erhebungsverfahren und einen Teil zu Lasten der Reaktionen auf inhaltliche Unterschiede zwischen den Variablen, die verschiedenen Konzepten zugeordnet sind. Damit wird es notwendig, bei der Prüfung der konvergenten und diskriminierenden Validität sowohl mehr als ein Konstrukt zu untersuchen als auch mehr als eine Erhebungsmethode zur Erfassung des gleichen Konstruktes zu benutzen, und wiederholte Messungen an den gleichen Vpn sind in der Regel erforderlich.

Die Aufbereitung der Daten geschieht übersichtlich in einer multitrait-multimethod Matrix, in der Korrelationen nach dem in Tab. 4 benutzten Schema angeordnet werden können. Die in der Hauptdiagonalen stehenden, eingeklammerten Werte stellen Reliabilitätskoeffizienten dar. Jeder dieser Koeffizienten bezieht sich auf eine Methode und ein Konzept (= „*monotrait-monomethod value*“). Die sich anschließenden Dreiecke mit den durchgezogenen Linien - „*heterotrait-monomethod triangle*“ - ergeben zusammen mit den zugehörigen Werten der Reliabilitätsdiagonale einen Einmethodenblock (*monomethod block*). Zwischen den gestrichelten Dreiecken befinden sich „Validitätsdiagonalen“, darin stellt jeder Koeffizient einen „*monotrait-heteromethod*“ Wert dar. Zwei angrenzende, gestrichelte Dreiecke, die mit möglicherweise verschiedenen Werten besetzt sein können und als „*heterotrait-heteromethod*“

triangles“ bezeichnet werden, bilden zusammen mit der zugehörigen Validitätsdiagonale einen Mehrmethodenblock.

Tabelle 4: Fiktive Korrelationen in einer multitrait-multimethod Matrix (nach Campbell & Fiske, 1959)

		Methode I			Methode II			Methode III		
<i>Konzepte</i>		A ₁	B ₁	C ₁	A ₂	B ₂	C ₂	A ₃	B ₃	C ₃
Methode I	A ₁	(.89)								
	B ₁	.51	(.89)							
	C ₁	.38	.37	(.76)						
Methode II	A ₂	.57	.22	.09	(.93)					
	B ₂	.22	.57	.10	.68	(.94)				
	C ₂	.11	.11	.46	.59	.58	(.84)			
Methode III	A ₃	.56	.22	.11	.67	.42	.33	(.94)		
	B ₃	.23	.58	.12	.43	.66	.34	.67	(.92)	
	C ₃	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

Damit Konstruktvalidität als nachgewiesen gelten kann, sollten folgende Forderungen erfüllt sein:

- (1) Die Koeffizienten in den Validitätsdiagonalen sollten signifikant von Null verschieden und substantiell sein; dies rechtfertigt den Schluß auf konvergierende Validität.
- (2) Ein Wert in der Validitätsdiagonale sollte größer sein als die Koeffizienten in den Zeilen und Spalten der gleichen heterotrait-heteromethod Dreiecke. D.h., die Validität der Variablen sollte größer sein als die Korrelation dieser Variablen mit irgendeiner anderen Variablen, die mit ihr weder das Konzept noch die Meßmethode gemeinsam hat (im Beispiel: $A_1A_2 = .57$ sollte größer sein als $A_1B_2 = .22$, $A_1C_2 = .11$, $A_1B_3 = .23$, $A_1C_3 = .11$ sowie $A_2B_1 = .22$, $A_2C_1 = .09$, $A_2B_3 = .43$, $A_2C_3 = .34$).
- (3) Eine Variable sollte höher mit einem unabhängigen Meßversuch korrelieren, der sich auf das gleiche Konzept bezieht, als mit Messungen, die sich „zufällig“ der gleichen Methode bedienen, jedoch zur Erfassung eines anderen Meßobjektes durchgeführt werden. Verglichen werden also Koeffizienten in der Validitätsdiagonale mit den Korrelationen in den zugehörigen heterotrait-monomethod Dreiecken (z.B. einerseits $A_1A_2 = .57$, andererseits $A_1B_1 = .51$; einerseits $A_1A_3 = .56$, andererseits $A_1C_1 = .38$. Für A₁ ist im Beispiel Kriterium 3 annähernd erfüllt, nicht jedoch z.B. für A₂).

- (4) Das gleiche Muster der korrelativen Beziehungen zwischen den Meßobjekten sollte sich in den heterotrait-Dreiecken sowohl der monomethod als auch der heteromethod Blöcke zeigen; die Beziehungen zwischen den Variablen sollten methoden-invariant sein (dies ist im Beispiel der Fall, trotz deutlicher Unterschiede in der absoluten Stärke der Korrelationen). Die Kriterien (2) bis (4) gestatten den Schluß auf die diskriminierende Validität.

Die grundlegende Auffassung der Validität, die den verschiedenen Ansätzen zu ihrer Bestimmung im Vorschlag von Campbell & Fiske in Abhebung zur Reliabilität gemeinsam ist, ist die *Übereinstimmung der Ergebnisse von unabhängigen Meßprozeduren am gleichen Objekt*. Die Beziehung zwischen den Konzepten Reliabilität und Validität erscheint aus dieser Perspektive über die Ähnlichkeit der Meßverfahren gegeben. Reliabilität ist die Übereinstimmung zwischen den Ergebnissen zweier Versuche, das gleiche Meßobjekt durch maximal ähnliche Verfahren zu erfassen. Validität drückt sich in der Übereinstimmung zwischen den Ergebnissen zweier Versuche aus, das gleiche Meßobjekt durch maximal verschiedene Methoden zu erfassen. Die Ähnlichkeit von Methoden kann man als variierend zwischen den Enden „maximal“ und „minimal“ eines Kontinuums auffassen; je stärker wir uns dem Pol maximale Ähnlichkeit nähern, desto stärker muß die Information als auf Reliabilität bezogen interpretiert werden. In diesem Sinne deuten Campbell & Fiske den Split-half Koeffizienten als in stärkerem Maße validitätsrelevant als ein Test-Retest-Koeffizient, da einzelne Items nicht völlig identisch sein können.

Es liegt nahe, Zeilen und Spalten der multitrait-multimethod Matrix inhaltlich anders als Campbell & Fiske zu definieren, wenn dies durch die Problemstellung erforderlich wird. In einer Untersuchung inhaltlich unterschiedlicher Erlebensdimensionen in einem intraindividuellen Konflikt verwendete Feger (1971; 1978, Kap. 2) für jedes der drei Konzepte drei verschieden formulierte Schätzskalen, insgesamt also neun Skalen. Der Versuchsplan entspricht somit einem design mit nested effects:

Konzept	A	B	C
Skala	1, 2, 3	4, 5, 6	7, 8, 9

Centra (1971) setzt an die Stelle unabhängiger Methoden diskrete soziale Gruppen; verschiedene Konzepte werden realisiert als Skalen, die Gruppenreaktionen auf verschiedene Aspekte eines multidimensionalen Wahrnehmungsraumes erfassen sollen. Ziel ist somit eine Einschätzung der Übereinstimmung der Gruppen; je ähnlicher die Urteile bei inhaltlich gleichen Skalen, desto höher die Validität, wobei der Gedanke im Hintergrund steht: „Wenn . . . diskrete Gruppen in gleicher Weise auf Skalen reagieren, die ihre Umwelt erfassen sollen, dann wird die Annahme einleuchtender, daß die Skalen Charakteristika oder Bedingungen dieser Umwelt wiedergeben.“

Krause (1972) hat die logischen Implikationen der konvergierenden und diskriminierenden Validierung untersucht. Er betont, daß andere Gründe als die Erfassung des gleichen Merkmals zur Konvergenz verschiedener Methoden führen können, und zeigt, daß sehr spezifische Bedingungen erfüllt sein müssen, damit nach den Kriterien von Campbell & Fiske auf Validität geschlossen werden kann. Er führt u.a. aus, die Meßverfahren müßten dafür bereits ein Mindestmaß an a priori-Validität besitzen, läßt aber unklar, was mit a priori-Validität gemeint ist und wie man nun ihr Vorliegen prüfen kann.

6.2 Neuere Entwicklungen zur Analyse von multitrait-multimethod Matrizen

Die Kritik, die an Campbell & Fiskes Ansatz geübt wurde, betont, die Entscheidung, ob die vier Kriterien oder auch jedes einzeln erfüllt sei, ließe sich im konkreten Fall nicht eindeutig treffen, weil die Kriterien nur vage verbal und nicht formal, exakt prüfbar definiert seien. Diese Kritik greifen Hubert & Baker (1978) auf und definieren die nur leicht modifizierten Kriterien formal, und zwar so, daß man inferenzstatistisch prüfen kann, ob sie erfüllt sind. Die gewählte Inferenzstatistik ist nonparametrisch, und man kann sie auch auf Korrelationskoeffizienten anwenden, die lediglich Ordinalskalenniveau in den Daten voraussetzen, wie etwa τ oder γ , sofern man τ oder γ selbst Intervallskalenniveau zuerkennt. Die später zu beschreibenden Weiterentwicklungen setzen durchweg Intervallskalenniveau der Daten voraus und gehen von Produkt-Moment-Korrelationskoeffizienten aus. Wie schwerwiegende Folgen Verstöße gegen diese impliziten Annahmen haben, ist nicht bekannt.

Mit anderen Argumenten hat Krause (1972, p. 183) infrage gestellt, ob Produkt-Moment-Korrelationen die geeigneten Koeffizienten seien, um Konvergenz und Diskriminanz von Verfahren festzustellen: „Whether or not two methods can be valid for the same trait is a question of codimensionality rather than (linear) prediction, and differences between methods in measurement distributions over the same n subjects can yield low correlations even when the pair of measurement sets is perfectly ordinally consistent, i.e., shows perfect Scalogram reproducibility, which is the necessary topological condition for codimensionality.“

Hubert & Baker definieren vier Indizes: (1) den Durchschnitt der Korrelationen zwischen den gleichen Merkmalen, (2) die Differenz zwischen dem ersten Index und dem Durchschnitt der Korrelationen zwischen verschiedenen Merkmalen, gemessen mit verschiedenen Methoden, (3) die Differenz zwischen dem ersten Index und dem Durchschnitt der Korrelationen zwischen gleichen Methoden, und (4) einen Index, der anzeigt, wie gleichartig das Korrelationsmuster der Merkmale über verschiedene Methoden hinweg ist. Um

die Bedeutung der numerischen Werte für die verschiedenen Indizes zu beurteilen, entwickeln Hubert & Baker eine Stichprobentheorie für die Zufallsverteilung der Indizes. Die allgemeine Nullhypothese besagt, daß die Korrelationen nicht die Struktur aufweisen, die einen Schluß auf konvergente und diskriminante Validität rechtfertigen könnte, daß die Zuordnung eines Meßverfahrens zu einer bestimmten Kombination von Methode und Merkmal eine zufällige ist. Die Indizes werden dann gegen die Zufallsverteilung auf einem vom Forscher gewählten Konfidenzniveau verglichen, und es lassen sich objektive Entscheidungen darüber fallen, welche Kriterien erfüllt sind.

Die erste kritische Weiterentwicklung des MTMM-Ansatzes stammt von Campbell & O'Connell (1967). Sie untersuchen die Frage, wie die Varianzquellen zusammenwirken, um die Höhe einer Korrelation in einer MTMM zu bestimmen. Die beiden angenommenen Varianzquellen sind (1) der Zusammenhang zwischen den Merkmalen und (2) der Zusammenhang zwischen den Meßmethoden. Für die untersuchten Datensätze mußte die Annahme eines additiven Zusammenwirkens der Varianzquellen zugunsten eines multiplikativen aufgegeben werden: Je höher eine heterotrait-heteromethod Korrelation ausfällt, desto stärker steigt sie dadurch an, daß die gleichen Variablen mit der gleichen Erhebungsmethode erfaßt werden. Als Konsequenz ergibt sich die Forderung, spezifische faktoranalytische Modelle zu prüfen, die nicht von einem additiven Zusammenwirken von Methoden- und Merkmalsfaktoren ausgehen.

Jackson (1969) kann als der eigentliche Begründer des faktoranalytischen Auswertungsansatzes von MTMM-Daten angesehen werden. Um das Vorgehen von Campbell & Fiske zu verbessern, geht er von folgenden Überlegungen aus: (1) Korrelationen sind in ihrer Höhe auch abhängig von Stichproben- und Meßfehler, was zum Fehlurteil führen kann, diskriminante Validität liege nicht vor. Daher sollte ein ideales Auswertungsverfahren diese Fehlerquellen, die Unreliabilität der Variablen und Korrelationen berücksichtigen. (2) Die Replizierbarkeit von Ergebnissen aus Validitätsstudien läßt sich vergrößern und Effekte von Stichprobenfehlern lassen sich mindern, wenn nicht einzelne Korrelationen, sondern das gesamte Korrelationsmuster der MTMM der Analyse zugrundegelegt wird. (3) Man kann nicht, wie Campbell & Fiske, generell davon ausgehen, daß heteromethod-Validitätskoeffizienten höher ausfallen als monomethod-heterotrait-Korrelationen. Das Analyseverfahren sollte die Beziehung zwischen Methodenvarianz und Merkmalsvarianz spezifizieren. (4) Zu welchen Schlüssen man über die konvergente und diskriminante Validität von Beobachtungen kommt, hängt von der Anzahl und relativen Ähnlichkeit der untersuchten Merkmale und Methoden ab. Wie Reliabilitätsaussagen auf den jeweiligen Erhebungsplan und die verwendete Stichprobe relativiert werden müssen, so sind Aussagen zur Konstruktvalidität darüber hinaus zu beziehen auf die mituntersuchten Merkmale und Methoden.

Jacksons eigenes Vorgehen, „multimethod factor analysis“ berücksichtigt nur Korrelationen zwischen Merkmalen, die mit verschiedenen Methoden erfaßt wurden, und gestattet nicht, Methodenfaktoren zu identifizieren. Außerdem können bei dieser Methode mathematisch-statistische Probleme auftreten (s. Schmitt et al. 1977, Conger, 1971). Jackson (1975) hat deshalb seinen Ansatz neu formuliert, wobei er an die Hauptkomponentenanalyse von Golding & Seidman (1974) anknüpfte. Die verschiedenen Verfahren werden von Schmitt et al. (1977) verglichen; auf den Ansatz von Boruch & Wolins (1970), der neben einem hypothetischen allgemeinen Faktor je einen pro Merkmal und pro Methode vorsieht, sei lediglich verwiesen. Hoffman & Tucker (1964), später z.B. auch Schmitt et al. (1977) haben Tuckers (1966) dreimodale Faktoranalyse zur Validitätsprüfung angewendet. Abgesehen von technischen und interpretativen Problemen, welche dieser Ansatz bietet, betont Jackson (1969, S. 35), diese Methode versuche, eine gemeinsame Faktorenstruktur der Merkmale zu identifizieren, die in mehr als einer Meßmethode wiederkehrt, und das Ziel sei nicht, für ein einzelnes Erhebungsverfahren oder ein bestimmtes Merkmal konvergierende und diskriminierende Validität zu untersuchen.

Das *pfadanalytische Vorgehen* bei Validierungsstudien haben meines Wissen Werts & Linn (1970) in die Psychologie eingeführt. Als Anwendungsbeispiele seien die Arbeiten von Kalleberg & Kluegel (1975), Ray & Heeler (1975) sowie Schmitt (1978) erwähnt. Die Pfadanalyse (zur Einführung: Heise, 1975) stellt ein System dar, mit dessen Hilfe Korrelationen in bezug auf ein vorgegebenes theoretisches Modell der Beziehungen zwischen den untersuchten Variablen interpretiert werden können. Campbell & Fiskes Kriterien werden im pfadanalytischen Modell im allgemeinen als die Hypothese interpretiert, jede Beobachtung für eine MTMM komme als Wirkung einer latenten Methodenvariablen und einer latenten Merkmalsvariablen zustande, plus einer Fehlerkomponenten. Methoden und Merkmale werden also als latente, nicht direkt gemessene, sondern aus verschiedenen Beobachtungen erschlossene Variablen aufgefaßt, wobei (zunächst) pro Methode und pro Merkmal eine spezifische latente Variable oder Faktor veranschlagt wird, weshalb dieses Vorgehen auch als „konfirmatorische Faktoranalyse“ bezeichnet wird. Das pfadanalytische Modell muß weiter die Beziehung zwischen Beobachtungen, Methoden- und Merkmalsfaktoren spezifizieren, z.B. festlegen, ob Methodenfaktoren als unabhängig von Merkmalsfaktoren gedacht werden, wie Campbell & Fiske annehmen. Die Auswertung geschieht oft mit Verfahren, die durch Jöreskog entwickelt wurden (Jöreskog 1969, 1970; Jöreskog & Sörbom, 1978). Ob Modell und Daten übereinstimmen, läßt sich inferenzstatistisch prüfen, auch verschiedene Modelle für die gleichen Beobachtungen lassen sich so vergleichen. Eine wichtige Validitätsfrage, ob bestimmte Methoden verschiedene Merkmale unterschiedlich verzerren, läßt sich für alle Methoden und Merkmale beantworten. Schmitt et al. (1977, S. 460) nennen folgende Vorteile einer pfadanalytischen Validierungsstudie: (1) sie zwingt zu einer exakten Formulie-

rung der Annahmen und deren logische Konsequenzen, (2) sie erlaubt, die Korrelationen zwischen Merkmalen, zwischen Methoden sowie zwischen Merkmalen und Methoden zu schätzen, (3) sie gestattet für jede Beobachtung die Einwirkung der Methoden- und Merkmalsfaktoren sowie der spezifischen Varianz abzuschätzen (Weiterentwicklung bei Bien 1980).

Man kann, wie Costner, ausdrücklich betonen, daß Beobachtungen nur den Charakter und die Funktion von *Indikatoren* für erschlossene, latente Variablen haben und explizit eine „Hilfstheorie“ (Costner, 1969) formulieren, deren Aufgabe es ist, die Beziehungen zwischen empirischen Indikatoren und Konstrukten zu beschreiben, und so die Grundlage für Urteile über die Güte der einzelnen Indikatoren bereitzustellen. Wichtig ist dabei auch, daß die inhaltliche Theorie nicht als an die empirische Prüfung durch nur eine spezielle Auswahl von Beobachtungen gebunden erscheint. Vielmehr sind im allgemeinen mehrere Hilfstheorien für eine inhaltliche Theorie möglich. Für Validitätsstudien haben Althausen & Heberlein (1970) sowie Costner & Schoenberg (1973) diese Zugangsweise vorgeschlagen, die sich mit konfirmatorischen Faktoranalysen verbinden läßt.

Insbesondere richtet sich in diesem Ansatz die Aufmerksamkeit auf die Frage, ob Meßfehler beim Erfassen des einen Konstrukts (oder verbunden mit einem bestimmten Indikator) mit Meßfehlern bei anderen Konstrukten oder Indikatoren korrelieren, ob also „differential bias“, „systematic measurement error“ oder „correlated measurement error“ vorliegt (Avison, 1978). Campbell & Fiske schließen auf konvergente Validität, wenn die monotrait-heteromethod Korrelationen hoch sind. Aus der Sicht der Indikator-Analyse impliziert diese Annahme zwei weitere: (1) die „epistemischen Koeffizienten“, die den Zusammenhang zwischen Indikator und Konstrukt beschreiben, sind substantiell hoch, (2) die Erhebungsmethoden sind unkorreliert. Ohne die zweite Annahme könnten hohe monotrait-heteromethod Korrelationen auf Zusammenhänge zwischen den Methoden zurückzuführen sein. Auf weitere, oft unrealistische Annahmen, die Campbell & Fiskes Kriterien für diskriminante Validität zugrunde liegen, machen Althausen & Heberlein, Althausen et al. (1971) sowie Avison (S. 438f.) aufmerksam. Althausen & Heberlein entwickeln eine eigene Vorgehensweise, die im Vergleich verschiedener Modelle für eine MTMM besteht. Costner & Schoenbergs Verfahren bewahrt den Forscher nicht nur vor falschen Schlüssen über den kausalen Zusammenhang von Variablen, der in Wirklichkeit auf korrelierten Meßfehlern beruht. Es erlaubt auch eine Diagnose von Schwächen bestimmter Indikatoren und eine genaue Abschätzung sowohl der epistemischen Koeffizienten als auch der „wahren“ Zusammenhänge zwischen den Konstrukten (s. auch Alwin 1974, Althausen 1974, Feger 1978, S. 77f.).

Die Informationen in den Beobachtungen einer Validitätsstudie lassen sich auch *varianzanalytisch* betrachten, wie dies Stanley (1961) vorgeschlagen hat.

Tabelle 5: Vergleich von Methoden für Studien über konvergierende und diskriminierende Validität
(nach Schmitt et al. 1977, S. 475).

ANALYSE-METHODE						
ZIELE	Campbell & Fiske	Stanley (ANOVA)	Pfadanalyse	Golding & Seidman	Jackson (1975)	Tucker (dreimodale FA)
1. Trennung von Merkmals- und Methodenvarianz.	NEIN	NEIN	Bei genügend großem N können verschiedene Modelle für Varianzquellen geprüft werden	Gemeinsame und spezifische Varianz sind in der ersten Transformation konfundiert	Gemeinsame und spezifische Varianz sind in der ersten Transformation konfundiert	Lösung abhängig von Komunalitäten-Schätzung und Anzahl der rotierten Faktoren
2. Evaluationskriterien für einzelne Merkmale und Methoden statt für Faktoren oder Komponenten	JA	NEIN, es gibt nur eine Gesamtaussage für die untersuchten Merkmale und Methoden	JA, jede Beobachtung wird als Resultat von Merkmals-Methoden- und spezifischer Varianz postuliert	NEIN, Konvergenz der Merkmalsfaktoren über Methoden ist das einzige Ziel	Konvergenz und Diskrimination geschehen auf dem Niveau der Faktorscores; Rotation gezielt auf einzelne Variablen	Ziel ist Schätzung der Interkorrelationen von Personen-, Methoden- und Merkmalsfaktoren

3. Basis für Bewertung des Ausmaßes an konvergenter und diskriminanter Validität	Prozentsatz, mit dem die Beziehungen zwischen Korrelationen den Kriterien entsprechen	Signifikanztest für Konvergenz von Merkmal und Methode	Chi ² -Tests für die Anpassungsgüte der angenommenen Modelle	Auftauchen eindeutiger Merkmalsfaktoren im zweiten Analyseschritt	Auftauchen von Merkmals- und Methodenfaktoren im zweiten Analyseschritt	Höhe der Korrelationen zwischen Personen-, Methoden- und Merkmalsfaktoren in der Kernmatrix
4. Bestimmung der Position der Merkmalsträger auf einem zugrundeliegenden Merkmal	NEIN	NEIN	JA	JA	JA	JA

Benutzt wird eine dreifaktorielle Varianzanalyse mit den systematischen Varianzquellen Methoden, Merkmale und Personen bzw. Merkmalsträger (Beispiele und Erweiterungen: Boruch et al. 1970, Kavanagh et al. 1971). Wie in den zuvor geschilderten Ansätzen liegt auch hier das allgemeine lineare Modell zugrunde, allerdings wird der Meßfehler nicht thematisiert. Zwar wird eine Validitätsaussage über alle verwendeten Methoden und Merkmale zugleich angestrebt, nicht jedoch für eine beliebige, bestimmte Erhebungsmethode für ein spezifisches Merkmal. Aus der Größe der Varianz zulasten der Merkmalsträger wird auf konvergierende Validität geschlossen. Dies ist, wie Schmitt et al. (1977) zu Recht betonen, nicht konvergierende Validität im Sinne von Campbell & Fiske, sondern das Ausmaß, in dem ein allgemeiner Faktor die gesamte MTMM zu erklären vermag. Insbesondere aus der Varianz zulasten der Interaktionen von Personen und Merkmalen sowie zwischen Personen und Methoden wird auf die diskriminierende Validität geschlossen.

Schmitt et al. (1977) haben verschiedene Methoden verglichen, mit deren Hilfe man konvergierende und diskriminierende Validität beurteilen kann. In ihrer Tab. 16 geben sie dazu eine Übersicht, zugleich die verschiedenen Ziele, die man mit einer Validitätsstudie verfolgen kann. Mit JA oder NEIN notieren sie, ob sich ein bestimmtes Ziel durch die betrachtete Methode erreichen läßt. Wir geben die übersetzte Tabelle als Tab. 5 wieder.

Literatur

Mit * gekennzeichnete Angaben werden nicht im Text zitiert, sind jedoch als interessante Quellen angeführt.

- Adams, R. S. 1970. Duration and incident frequency as observation indices. *Educational and Psychological Measurement*, 30, 669-674.
- Althauser, R. P. 1974. Inferring validity from the multitrait-multimethod matrix: Another assessment. In: Costner, H. L. (ed.): *Sociological Methodology*, 1973/74. San Francisco: Jossey-Bass, 106-127.
- Althauser, R. P. & Heberlein, T. A. 1970. Validity and the multitrait-multimethod matrix. In: Borgatta E. F. & Bohrnstedt, G. W. (eds): *Sociological methodology* 1970. San Francisco: Jossey-Bass, 151-169.
- Althauser, R. P., Heberlein, T. A. & Scott, R. A. 1971. A causal assessment of validity: The augmented multitrait-multimethod matrix. In: Blalock, H. M. (ed.): *Causal models in the social sciences*. London: Macmillan Press.
- Alwin, D. F. 1974. Approaches to the interpretation of relationships in the multitrait-multimethod matrix. In: Costner, H. L. (ed.): *Sociological Methodology*, 1973/74. San Francisco: Jossey-Bass, 79-105.
- Amerine, M. A. & Roessler, E. B. 1976. *Wines: Their sensory evaluation*. San Francisco: Freeman & Co.

- Arrington, R. E. 1939. Time sampling studies of child behavior. *Psychological Monographs*, 51, 2.
- Arrington, R. E. 1943. Time sampling studies of social behavior: a critical review of techniques and results with research suggestions. *Psychological Bulletin*, 40, 81-124.
- Asendorpf, J. & Wallbott, H. G. 1979. Maße der Beobachterübereinstimmung: Ein systematischer Vergleich. *Zeitschrift für Sozialpsychologie*, 10, 243-252.
- Assakul, K. & Procter, C. H. 1967. Testing independence in two-way contingency tables with data subject to misclassification. *Psychometrika*, 32, 67-76.
- Avison, W. R. 1978. Auxiliary theory and multitrait-multimethod validation: A review of two approaches. *Applied Psychological Measurement*, 2, 431-447.
- Bales, R. F. 1950. *Interaction process analysis*. Cambridge, Mass.: Addison-Wesley.
- Bales, R. F. 1968. Die Interaktionsanalyse: Ein Beobachtungsverfahren zur Beobachtung kleiner Gruppen. In: König, R. (Hrsg.): *Beobachtung und Experiment in der Sozialforschung*. Köln: Kiepenheuer & Witsch, 6. Aufl., 148-170.
- Barker, R. 1951. *One boy's day*. New York. Harper & Row.
- Bartko, J. J. 1976. On various intraclass correlation reliability coefficients. *Psychological Bulletin*, 83, 762-765.
- Bergan, J. R. 1980. Measuring observer agreement using the quasi-independence concept. *Journal of Educational Measurement*, 17, 59-69.
- Bien, W. 1980. Ein Vergleich von Datenerhebungsmethoden zu kognitiven sozialen Strukturen durch die Analyse einer Multitrait-Multimethod-Matrix. Dissertation, RWTH Aachen.
- Bishop, Y. M., Fienberg, S. & Holland, P. 1974. *Discrete multivariate analysis: Theory and practice*. Boston: M. I. T. Press.
- Blalock, H. M. Jr. 1969. Multiple indicators and the causal approach to measurement error. *American Journal of Sociology*, 75, 264-272.
- Blalock, H. M. Jr. 1970. Estimating measurement error using multiple indicators and several points in time. *American Sociological Review*, 35, 101-111.
- Bock, R. D. 1956. The selection of judges for preference testing. *Psychometrika*, 21, 349-366.
- Böltkén, F. 1976. *Auswahlverfahren. Eine Einführung für Sozialwissenschaftler*. Stuttgart: Teubner (Studienskripten).
- Bohrnstedt, G. W. 1969. Observations on the measurement of change. In: Borgatta, E. F. (ed.): *Sociological Methodology 1969* San Francisco: Jossey-Bass, 113-133.
- Borg, I. 1977. Some basic concepts of facet theory. In: Lingoes, J. C. (ed.): *Geometric representations of relational data*. Ann Arbor, Mich.: Mathesis Press, 65-102.
- Boruch, R. F. & Wolins, L. 1970. A procedure for estimations of trait, method, and error variance attributable to a measure. *Educational and Psychological Measurement*, 30, 547-574.

- Boruch, R. F., Larkin, J. D., Wolins, L. & Mac Kinney, A. C. 1970. Alternative methods of analysis: Multitrait-multimethod data. *Educational and Psychological Measurement*, 30, 833-853.
- *Brannigan, C. R. & Humphries, D. A. 1972. Human non-verbal behavior, a means of communication. In: Blurton Jones, N. (ed.): *Ethological studies of child behavior*. Cambridge: University Press, 37-64.
- Bredenkamp, J. 1972. *Der Signifikanztest in der psychologischen Forschung*. Frankfurt: Akademische Verlagsgesellschaft.
- Bredenkamp, J. 1980. *Theorie und Planung psychologischer Experimente*. Darmstadt: Steinkopf.
- Brennan, R. L. & Light, R. J. 1974. Measuring agreement when two observers classify people into categories not defined in advance. *The British Journal of Mathematical and Statistical Psychology*, 27, 154-163.
- Bross, I. 1954. Misclassification in 2×2 tables. *Biometrics*, 10, 478-486.
- Caldwell, B. M. 1969. A new „approach“ to behavioural ecology. In: Hull, J. P. (ed.): *Minnesota Symposia on Child Psychology*, Vol. 2, Minneapolis: Univ. of Minnesota Press.
- Campbell, D. T. & Fiske, D. 1959. Convergent and discriminant Validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Campbell, D. T. & Stanley, J. C. 1963. Experimental and quasi-experimental designs for research in teaching. In: Gage, N. L. (ed.): *Handbook of research in teaching*. Chicago: Rand McNally, 171-246.
- Campbell, D. T. & O'Connell, E. J. 1967. Method factors in multimethod-multitrait matrices: Multiplicative rather than additive? *Multivariate Behavioral Research*, 2, 409-426.
- Canter, D. 1977. Is there a mapping sentence for architectural semiotics? Paper prepared for the Semiotics Symposium at Edra 8, Spring.
- Canter, D. 1977. Children in hospital: A facet theory approach to person/place synomorphy. *Journal of Architectural Research*, 6, 20-32 (b).
- Centra, J. A. 1971. Validation by the multigroup-multiscale matrix: an adaptation of Campbell and Fiske's convergent and discriminant validation procedure. *Educational and Psychological Measurement*, 31, 675-683.
- Cicchetti, D. V., Aivano, S. L. & Vitale, J. 1977. Computer programs for assessing rater agreement and rater bias for qualitative data. *Educational and Psychological Measurement*, 37, 195-201.
- Clarke, A. H. & Ellgring, J. H. 1978. Verfahren zur halbautomatischen Bearbeitung von Videoaufzeichnungen. In: Heimchen, H. & Rengardt, E. (Hrsg.): *Fernsehen in der Psychiatrie*. Stuttgart: Thieme, 107-115.
- Cleary, T. A. & Linn, R. L. 1969. Error of measurement and the power of a statistical test. *British Journal of Mathematical and Statistical Psychology*, 22, 49-55.
- Cochran, W. G. 1953. *Sampling techniques*. New York: Wiley.

- Cochran, W. G. 1968. Errors of measurement in statistics. *Technometrics*, 10, 637-666.
- Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- Cohen, J. 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.
- Cohen, J. 1969. *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Conger, A. J. 1971. An evaluation of multimethod factor analysis. *Psychological Bulletin*, 75, 416-420.
- Conger, A. J. 1980. Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, 88, 322-328.
- Costner, H. L. 1969. Theory, deduction, and rules of correspondence. *American Journal of Sociology*, 75, 245-263.
- Costner, H. L. & Schoenberg, R. 1973. Diagnosing indicator ills in multiple indicator models. In: Goldberger, A. S. & Duncan, O. D. (eds): *Structural equation models in the social sciences*. New York: Seminar Press, 167-199.
- Cranach von, M. & Frenz, H. G. 1969. Systematische Beobachtung. In: Graumann, C. F. (Hrsg.): *Handbuch der Psychologie*, Band 7/1, Sozialpsychologie. Göttingen: Hogrefe, 269-331.
- Cronbach, L. J. & Meehl, P. E. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Cronbach, L. J., Rajaratnam, N. & Gleser, G. C. 1963. Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology*, 16, 137-163.
- Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. 1972. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972.
- Cureton, E. E. 1958. The average Spearman rank criterion correlation when ties are present. *Psychometrika*, 23, 271-272.
- Cureton, E. E. 1965. The average Spearman rank correlation when ties are present: A correlation. *Psychometrika*, 30, 377.
- Davis, F. B. 1974. *Standards for educational and psychological tests*. Washington: American Psychological Association.
- Deming, W. E. 1950. *Some theory of sampling*. New York: Wiley.
- Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26, 297-302.
- *Dickman, H. R. 1963. The perception of behavioral units. In: Barker, R. G. (ed.): *The stream of behavior*. New York: Appleton-Century-Crofts.
- Duncan, O. D. 1969. Some linear models for two-wave, two-variable panel analysis. *Psychological Bulletin*, 72, 177-182.

- Duncan, O. D. 1972. Unmeasured variables in linear models for panel analysis. In: Costner, H. L. (ed.): *Sociological Methodology 1972*. San Francisco: Jossey-Bass, 36-82.
- Duncan, O. D. 1975. Some linear models for two-wave, two variable panel analysis, with one-way causation and measurement error. In: Blalock, H. M. Jr., Aganegian, A., Borodkin, F. N., Bondon, R. & Capecch, V. (eds): *Quantitative Sociology: International perspectives on mathematical and statistical modeling*. New York: Academic Press, 285-306.
- Duncan, S. 1969. Nonverbal communication. *Psychological Bulletin*, 72, 118-137.
- Duncan, S. Jr. & Fiske, D. W. 1977. *Face-to-face interaction: Research, methods, and theory*. Hillsdale: Erlbaum.
- Ebel, R. L. 1951. Estimation of the reliability of ratings. *Psychometrika*, 16, 407-424.
- Everitt, B. S. 1968. Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 21, 97-103.
- Faßnacht, G. 1979. *Systematische Verhaltensbeobachtung*. München: Reinhardt (UTB Nr. 889).
- Feger, H. 1971. Probleme der Reliabilität und einer konvergierenden und diskriminierenden Validierung von Selbstbeobachtungsdaten. *Archiv für Psychologie*, 123, 1-16.
- Feger, H. 1978. *Konflikterleben und Konfliktverhalten*. Bern: Huber.
- Feger, H. 1979. Einstellungsstruktur und Einstellungsänderung: Ergebnisse, Probleme und ein Komponentenmodell der Einstellungsobjekte. *Zeitschrift für Sozialpsychologie*, 10, 331-349.
- Finn, R. H. 1970. A note on estimating the reliability of categorical data. *Educational and Psychological Measurement*, 30, 71-76.
- Finn, R. H. 1972. Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement*, 3.5, 255-265.
- Fischer, G. W. 1976. Multidimensional Utility models for risky and riskless choice. *Organizational Behavior and Human Performance*, 17, 127-146.
- Fishbein, M. & Ajzen, I. 1975. *Belief, attitude, intention, and behavior: An introduction to theory and research*. Reading, Mass.: Addison-Wesley.
- Fleiss, J. L. 1966. Assessing the accuracy of multivariate observations. *Journal of the American Statistical Association*, 61, 403-412.
- Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382.
- Fleiss, J. L. 1973. *Statistical methods for rates and proportions*. New York: Wiley.
- Fleiss, J. L. 1975. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, 31, 651-659.

- Fleiss, J. L. & Cohen, J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33, 613-619.
- Fleiss, J. L. & ShROUT, P. E. 1978. Approximate interval estimation for a certain intraclass correlation coefficient. *Psychometrika*, 43, 259-262.
- Fleiss, J. L., Cohen, J. & Everitt, B. S. 1969. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72, 323-327.
- Fleiss, J. L., Nee, J. C. M. & Landis, J. R. 1979. Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, 86, 974-977.
- Fleiss, J. L., Spitzer, R. L., Endicott, J. & Cohen, J. 1972. Quantification of agreement in multiple psychiatric diagnosis. *Archives of General Psychiatry*, 26, 168-171.
- Frey, S. & Pool, J. 1976. A new approach to the analysis of visible behavior. Forschungsbericht aus dem Psychologischen Institut der Universität Bern.
- Frick, R. & Semmel, M. I. 1978. Observer agreement and reliabilities of classroom observational measures. *Review of Educational Research*, 48, 157-184.
- Friedrichs, J. & Lüdtkke, H. 1973. *Teilnehmende Beobachtung*. Beltz, Weinheim.
- Gleser, G. C., Cronbach, L. J. & Rajaratnam, N. 1965. Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 30, 395-418.
- Golding, S. L. & Seidman, E. 1974. Analysis of multitrait-multimethod matrices: A two step principal components procedure. *Multivariate Behavioral Research*, 9, 479-496.
- Goodman, L. A. & Kruskal, W. H. 1954. Measures of association for cross classification. *Journal of the American Statistical Association*, 49, 732-764.
- Goodman, L. A. & Kruskal, W. H. 1959. Measures of association for cross classification. II: Further discussion and references. *Journal of the American Statistical Association*, 54, 123-163.
- Goodman, L. A. & Kruskal, W. H. 1963. Measures of association for cross classification. III: Approximate sampling theory. *Journal of the American Statistical Association*, 58, 310-364.
- Goodman, L. A. & Kruskal, W. H. 1972. Measures of association for cross classification. IV: Simplification of asymptotic variances. *Journal of the American Statistical Association*, 67, 415-421.
- Graumann, C. F. 1966. Grundzüge der Verhaltensbeobachtung. In: Meyer, E. (Hrsg.): *Fernsehen in der Lehrerbildung*. München: Manz Verlag, 86-107.
- Green, D. M. & Swets, J. A. 1966. *Signal detection theory and psychophysics*. New York: Wiley.
- Grubbs, F. E. 1948. On estimating precision of measuring instruments and product variability. *Journal of the American Statistical Association*, 43, 243-264.
- Grubbs, F. E. 1973. Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics*, 15, 53-66.
- Grüner, K.-W. 1974. *Beobachtung*. Stuttgart. Teubner (Studienskripte).

- Guttman, L. 1959. A structural theory for intergroup beliefs and action. *American Sociological Review*, 24, 318-328.
- Guttman, L. 1971. Measurement as structural theory. *Psychometrika*, 36, 329-347.
- Guttman, L. & Guttman, R. 1976. A theory of behavioral generality and specificity during mild stress. *Behavioral Science*, 21, 469-477.
- Haggard, E. A. 1958. *Intraclass correlation and the analysis of variance*. New York: Dryden.
- Hannan, M. T., Robinson, R. & Warren, J. T. 1974. The causal approach to measurement error in panel analysis: Some further contingencies. In: Blalock, H. M. Jr. (ed.): *Measurement in the social sciences: Theories and strategies*. Chicago: Aldine, 293-323.
- Hansen, M. H., Hurwitz, W. N. & Madow, W. G. 1953. *Sample survey methods and theory*. New York: Wiley.
- Hasemann, K. 1964. Verhaltensbeobachtung. In: Heiss, R. (Hrsg.): *Handbuch der Psychologie*, Band 6, *Psychologische Diagnostik*. Göttingen: Hogrefe, 807-836.
- Hauser, R. M. & Goldberger, A. S. 1971. The treatment of unobservable variables in path analysis. In: Costner, H. L. (ed.): *Sociological Methodology 1971*. San Francisco: Jossey-Bass, 81-117.
- Hayes, D. P., Meltzer, L. & Wolf, G. 1970. Substantive conclusions are dependent upon techniques of measurement. *Behavioral Science*, 15, 265-268.
- Heise, D. R. 1969. Separating reliability and stability in test-retest correlation. *American Sociological Review*, 34, 93-101. (Nachdruck in: Blalock, H. M. (ed.): *Causal models in the social sciences*. London: MacMillan, 1971.)
- Heise, D. R. 1970. Causal inference from panel data. In: Borgatta, E. F. & Bohrnstedt, G. W. (eds): *Sociological Methodology 1970*. San Francisco: Jossey-Bass, 3-27.
- Heise, D. R. 1975. *Causal analysis*. New York: Wiley.
- Hendel, D. D. & Weiss, D. J. 1970. Individual inconsistency and reliability of measurement. *Educational and Psychological Measurement*, 30, 579-593.
- Heyns, R. W. & Zander, A. F. 1953. Observation of group behavior. In: Festinger, L. & Katz, D. (eds): *Research methods in the behavioral sciences*. New York: Dryden Press, 381-417.
- Hildebrand, D. K., Laing, J. D. & Rosenthal, H. 1977. *Analysis of ordinal data*. Beverly Hills: Sage Publications.
- Hollenbeck, A. R. 1978. Problems of reliability in observational research. In: Sackett, G. P. (ed.): *Observing behavior*. Vol. II. Baltimore: Univ. Park Press, 79-98.
- Holley, J. W. & Guilford, J. P. 1964. A note on the G index of agreement. *Educational and Psychological Measurement*, 24, 749-753.
- Holley, J. W. & Lienert, G. A. 1974. The G index of agreement in multiple ratings. *Educational and Psychological Measurement*, 34, 817-822.
- Huber, H. P. 1973. *Psychometrische Einzelfalldiagnostik*. Weinheim: Beltz.

- Hubert, L. 1977. Kappa revisited. *Psychological Bulletin*, 84, 289-297.
- Hubert, L. 1977. Nominal scale response agreement as a generalized correlation. *British Journal of Mathematical and Statistical Psychology*, (a), 30, 98-103.
- Hubert, L. J. 1979. Comparison of sequences. *Psychological Bulletin*, 86, 1098-1106.
- Hubert, L. J. 1979. Generalized concordance. *Psychometrika*, (a), 44, 135-142.
- Hubert, L. 1979. Matching models in the analysis of cross-classifications. *Psychometrika*, (b), 44, 21-41.
- Hubert, L. J. & Baker, F. B. 1978. Analyzing the multitrait-multimethod matrix. *Multivariate Behavioral Research*, 13, 163-179.
- Hubert, L. J. & Baker, F. B. 1978. Evaluating the conformity of sociometric measurements. *Psychometrika*, (a), 43, 31-41.
- *Hubert, L. J. & Baker, F. B. 1979. A note on analysing the multitrait-multimethod-matrix: An application of a generalized proximity function comparison. *British Journal of Mathematical and Statistical Psychology*, 32, 179-184.
- Huck, S. W. 1978. A modification of Hoyt's analysis of variance reliability estimation procedure. *Educational and Psychological Measurement*, 38, 725-736.
- Hunter, J. E. & Cohen, S. H. 1974. Correcting for unreliability in nonlinear models of attitude Change. *Psychometrika*, 39, 445-468.
- Hutt, S. J. & Hutt, C. 1974. *Direct Observation and measurement of behavior*. Springfield: Thomas, 2. Aufl.
- Jackson, D. N. 1969. Multimethod factor analysis in the evaluation of convergent and discriminant validity. *Psychological Bulletin*, 72, 30-49.
- Jackson, D. N. 1975. Multimethod factor analysis: A reformulation. *Multivariate Behavioral Research*, 10, 259-275.
- Jacobson, A. L. & Lalw, N. M. 1974. An empirical and algebraic analysis of alternative techniques for measuring unobserved variables. In: Blalock, H. M. Jr. (ed.): *Measurement in the social sciences*. Chicago: Aldine, 215-242.
- Jahoda, M., Deutsch, M. & Cook, S. W. 1968. Beobachtungsverfahren. In: König, R. (Hrsg.): *Beobachtung und Experiment in der Sozialforschung*. Köln: Kiepenheuer & Witsch, 6. Aufl., 77-96.
- Janson, S. & Vegelius, J. 1979. On generalizations of the G-index and the phi-coefficient to nominal scales. *Multivariate Behavioral Research*, 14, 255-269.
- Joe, G. W. & Woodward, J. A. 1976. Some developments in multivariate generalizability. *Psychometrika*, 41, 205-217.
- Jöreskog, K. G. 1969. A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34, 183-202.
- Jöreskog, K. G. 1970. A general method for analysis of covariance structures. *Biometrika*, 57, 239-251.
- Jöreskog, K. G. & Sörbom, D. 1976. Statistical models and methods for analysis of longitudinal data. In: Aigner, D. J. & Goldberger, A. S. (eds): *Latent variables in socioeconomic models*. Amsterdam: North Holland.

- Jöreskog, K. G. & Sörbom, D. 1978. LISREL - IV: A general computer program for estimation of linear structural equations by maximum likelihood methods. User's Guide. Chicago: International Educational Service.
- Jordan, J. E. 1971. Construction of a Guttman facet designed crosscultural attitude-behavior scale toward mental retardation. *American Journal of Mental Deficiency*, 76, 201-219.
- Kalbermatten, U. & v. Cranach, M. 1980. Hierarchisch aufgebaute Beobachtungssysteme zur Handlungsanalyse. In: Winkler, P. (Hrsg.): *Methoden zur Analyse von Face-to-Face Situationen*. Stuttgart: Metzler.
- Kalleberg, A. L. & Kluegel, J. R. 1975. Analysis of the multitrait-multimethod matrix: Some limitations and an alternative. *Journal of Applied Psychology*, 60, 1-9.
- Kaufman, I. C. & Rosenblum, L. A. 1966. A behavioral taxonomy for *M. nemestrina* and *M. radiata*; based on longitudinal observation of family groups in the laboratory. *Primates*, 7, 205-258.
- Kavanagh, M. J., MacKinney, A. C. & Wolins, L. 1971. Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin*, 75, 34-39.
- Kaye, K. 1980. Estimating false alarms and missed events from interobserver agreement: A rationale. *Psychological Bulletin*, 88, 458-468.
- Kendall, M. G. 1948. *Rank correlation methods*. London: Griffin.
- *Kenny, D. A. 1976. An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology*, 12, 247-252.
- Kent, R. N. & Foster, S. L. 1977. Direct observational procedures: Methodological issues in naturalistic settings. In: Ciminero, A. R., Calhoun, K. S. & Adams, H. E. (eds): *Handbook of behavioral assessment*. New York: Wiley, 279-328.
- Keys, A. & Kihlberg, J. K. 1963. Effect of misclassification on estimated relative prevalence of a characteristic. *American Journal of Public Health*, 53, 1656-1665.
- Kish, L. 1953. Selection of a Sample. In: Festinger, L. & Katz, D. (eds): *Research methods in the behavioral sciences*. London: Staples Press.
- Klauer, K. J. 1978. Kontentvalidität. In: Klauer, K. J. (Hrsg.): *Handbuch der pädagogischen Diagnostik*. Band 1. Düsseldorf: Schwann, 225-255.
- Kluckhohn, F. R. 1956. Die Methode der teilnehmenden Beobachtung. In: König, R. (Hrsg.): *Beobachtung und Experiment in der Sozialforschung*. Köln: Kiepenheuer & Witsch, (3. Aufl. 1965), 97-114.
- Koch, G. G. 1969. The effect of non-sampling errors on measures of association in 2 x 2 contingency tables. *Journal of the American Statistical Association*, 64, 852-863.
- König, R. 1962. Die Beobachtung. In: König, R. (Hrsg.): *Handbuch der empirischen Sozialforschung*, Band I. Stuttgart: Enke, 107-135.

- Kraemer, H. C. & Korner, A. F. 1976. Statistical alternatives in assessing reliability, consistency, and individual differences for quantitative measures: Application to behavior measures of neonates. *Psychological Bulletin*, 83, 914-921.
- Krause, M. S. 1972. The implications of convergent and discriminant validity data for instrument Validation. *Psychometrika*, 37, 179-186.
- Krippendorff, K. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, 30, 61-70.
- Krippendorff, K. 1970. Bivariate agreement coefficients for reliability of data. In: Borgatta, E. F. & Bohrnstedt, G. W. (eds): *Sociological methodology 1970*. San Francisco: Jossey-Bass (a).
- Kruglanski, A. W. 1975. The human subject in the psychological experiment: Fact and artifact. In: Berkowitz, L. (ed.): *Advances in experimental social psychology*, Vol. 8. New York: Academic Press, Vol. 8, 101-147.
- Kruskal, W. H. 1958. Ordinal measures of association. *Journal of the American Statistical Association*, 53, 814-861.
- Landis, J. R. & Koch, G. G. 1975. A review of statistical methods in the analysis of data arising from observer reliability studies. *Statistica Neerlandica*, 29, Part I: 101-123, Part II: 151-161.
- Landis, J. R. & Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Landis, J. R. & Koch, G. G. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, (a), 33, 363-374.
- Lawlis, G. F. & Lu, E. 1972. Judgment of counseling process: Reliability, agreement, and error. *Psychological Bulletin*, 78, 17-20.
- Leach, G. M. 1972. A comparison of the social behavior of some normal and problem children. In: Blurton Jones, N. (ed.): *Ethological studies of Child behavior*. Cambridge: Univ. Press, 249-281.
- Levy, P. 1974. Generalizability studies in clinical settings. *British Journal of Social and Clinical Psychology*, 13, 161-172.
- Levy, S. & Guttman, L. 1975. Structure and dynamics of worries. *Sociometry*, 38, 445-473.
- *Lienert, G. A. 1961. *Testaufbau und Testanalyse*. Weinheim: Beltz, 2. Aufl. 1967.
- Lienert, G. A. 1973. *Verteilungsfreie Methoden in der Biostatistik*. Band I. Meisenheim: Hain.
- Light, R. J. 1971. Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, 76, 365-377.
- Lisch, R. & Kriz, J. 1978. *Grundlagen und Modelle der Inhaltsanalyse*. Reinbek bei Hamburg: Rowohlt.

- Longabaugh, R. 1980. The systematic observation of behavior in naturalistic settings. In: Triandis, H. C. & Berry, J. W. (eds): *Handbook of Cross-cultural psychology*. Vol. 2: Methodology. Boston: Allyn & Baton, pp. 57-126.
- Lord, F. M. & Novick, M. R. 1968. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.
- Lu, K. H. 1971. Statistical control of „impurity“ in the estimation of test reliability. *Educational and Psychological Measurement*, 31, 641-655.
- Luce, R. D. 1963. Detection and recognition. In: Luce, R. D., Bush, R. R. & Galanter, E. (eds): *Handbook of mathematical psychology*. Vol. I. New York: Wiley, 103-189.
- Luce, R. D. & Galanter, E. 1963. Discrimination. In: Luce, R. D., Bush, R. R. & Galanter, E. (eds): *Handbook of mathematical psychology*. Vol. I. New York: Wiley, 191-243 (a).
- Luce, R. D. & Galanter, E. 1963. Psychophysical scaling. In: Luce, R. D., Bush, R. R. & Galanter, E. (eds): *Handbook of mathematical psychology*. Vol. I. New York: Wiley, 245-307 (b).
- Lyerly, S. B. 1952. The average Spearman rank correlation coefficient. *Psychometrika*, 17, 421-428.
- Manz, W. 1974. Beobachtung verbaler Kommunikation im Laboratorium. In: van Koolwijk, J. & Wicken-Mayser, M. (Hrsg.): *Techniken der empirischen Sozialforschung. Band III, Erhebungsmethoden: Beobachtung und Analyse von Kommunikation*. München: Oldenbourg, 27-65.
- Maxwell, A. E. & Pilliner, A. E. G. 1968. Deriving coefficients of reliability and agreement for ratings. *British Journal of Mathematical and Statistical Psychology*, 21, 105-116.
- Mayer, L. S. & Younger, M. S. 1974. Multiple indicators and the relationship between abstract variables. In: Heise, D. P. (ed.): *Sociological methodology 1975*. San Francisco: Jossey-Bass, 191-211.
- McCall, G. J. & Simons, J. L. 1969. *Issues in participant Observation: A text and reader*. Reading, Mass.: Addison-Wesley.
- McDermott, P. A. & Watkins, M. W. 1979. A Fortran program for testing agreement of multiple observers with a categorical standard on nominal scales. *Educational and Psychological Measurement*, 39, 669-672.
- McGrew, W. C. 1972. *An ethological study of children's behavior*. New York: Academic Press.
- McKinlay, S. M. 1975. The design and analysis of the observational study - A review. *Journal of the American Statistical Association*, 70, 503-520.
- McNicol, D. 1972. *A primer of signal detection theory*. London: Allen & Unwin.
- Medley, D. M. & Mitzel, H. E. 1970. Measuring classroom behavior by systematic Observation. In: Gage, N. L. (ed.): *Handbook of research on teaching*. Chicago: Rand McNally, 1963, 247-328. Deutsche Bearbeitung: Schulz, W., Teschner, W. P. & Voigt, J.: *Verhalten im Unterricht. Seine Erfassung durch Beobachtungs-*

- verfahren. In: Ingenkamp, K. (Hrsg.): Handbuch der Unterrichtsforschung, Teil I. Weinheim: Beltz, 632-851.
- Mees, U. & Selg, H. 1977. Verhaltensbeobachtung und Verhaltensmodifikation. Stuttgart: Klett.
- Mote, V. L. & Anderson, R. L. 1965. An investigation of the effect of misclassification on the properties of chi square tests in the analysis of categorical data. *Biometrika*, 52, 95-109.
- Munroe, R. L. 1973. Differential predictions based on time sampling versus event sampling behavior observations. Paper presented at Mathematical Social Science Board Conference on Human Behavior Observations. Monroeville, Pa. October.
- Nay, W. R. & Kerkhoff, T. 1974. Informational feedback in training behavioral coders. *Psychological Reports*, 35, 1175-1181.
- Naylor, J. C. & Schenck, E. A. 1966. Q_m as an „error free“ index of rater agreement. *Educational and Psychological Measurement*, 26, 815-824.
- Naylor, J. C., Dudycha, A. L. & Schenck, E. A. 1967. An empirical comparison of Q_α and Q_m as indices of rater policy agreement. *Educational and Psychological Measurement*, 27, 7-20.
- Newtonson, D. 1973. Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28, 28-38.
- Newtonson, D. 1976. Foundations of attribution: The unit of perception of ongoing behavior. In: Harvey, J., Ickes, W. & Kidd, R. (eds): *New directions in attribution research*. Hillsdale, N. J.: Erlbaum, 223-247.
- Newtonson, D. & Engquist, G. 1976. The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, 12, 436-450.
- Newtonson, D., Engquist, G. & Bois, J. 1977. The objective basis of behavior units. *Journal of Personality and Social Psychology*, 35, 847-862.
- Nicewander, W. A. & Price, J. M. 1978. Dependent variable reliability and the power of significance tests. *Psychological Bulletin*, 85, 405-409.
- Nußbaum, A. 1980. Konstruktion, Planung und Analyse lehrzielorientierter Tests auf der Grundlage der Generalisierbarkeitstheorie. Dissertation, RWTH Aachen.
- Overall, J. E. 1968. Estimating individual rater reliabilities from analysis of treatment effects. *Educational and Psychological Measurement*, 28, 255-264.
- Peak, H. 1953. Problems of objective observation. In: Festinger, L. & Katz, D. (eds): *Research methods in the behavioral sciences*. New York: Holt, Rinehart, and Winston, 243-299.
- Pearson, K. 1901. Mathematical contributions to the theory of evolution. IX On the principle of homotyposis and its relation to heredity, to variability of the individual, and to that of rate. Part I: Homotyposis in the vegetable kingdom *Philosophical Transactions of the Royal Society (London)*, Series A, 197, 385-479.
- Polansky, N., Freeman, W., Horowitz, M., Irwin, L., Paponia, N., Rapaport, D. & Whaley, F. 1949. Problems of interpersonal relations in research on groups. *Human Relation*.

- Rajaratnam, N. 1960. Reliability formtdas for independent decision data when reliability data are matched. *Psychometrika*, 25, 261-271.
- Ray, M. L. & Heeler, R. M. 1975. Analysis techniques for exploratory use of the multitrait-multimethod matrix. *Educational and Psychological Measurement*, 35, 255-265.
- Reid, J. B. 1970. Reliability assessment of observational data: a possible methodological problem. *Child Development*, 41, 1143-1150.
- Rock, D. A., Werts, C. E., Linn, R. L. & Jöreskog, K. G. 1977. A maximum likelihood Solution to the errors in variables and errors in equation model. *Multivariate Behavioral Research*, 12, 187-197.
- Rogot, E. 1961. A note on measurement errors and detecting real differences. *Journal of the American Statistical Association*, 56, 319-319.
- Romanczyk, R. G., Kent, R. N., Diamant, C. & O'Leary, K. D. 1973. Measuring the reliability of observational data: a reactive process. *Journal of Applied Behavior Analysis*, 6, 175-184.
- Rosenthal, R. 1965. The volunteer subject. *Human Relations*, 18, 389-406.
- Rosenthal, R. & Rosnow, R. L. (eds). 1969. *Artifact in behavioral research*. New York: Academic Press.
- Rosenthal, R. & Rosnow, R. L. 1975. *The volunteer subject*. New York: Wiley.
- Rudinger, G. & Feger, H. 1970. Die Beurteilung formaler Verhaltensmerkmale durch Rating-Skalen: Eine Generalisierbarkeitsstudie. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 2, 96-112.
- Sackett, G. P. 1978. Measurement in observational research. In: Sackett, G. P. (ed.): *Observing behavior*. Vol. II. Baltimore: Univ. Park Press, 25-43.
- Sackett, G. P., Ruppenthal, G. C. & Gluck, J. 1978. Introduction: An overview of methodological and statistical problems in observational research. In: Sackett, G. P. (ed.): *Observing behavior*. Vol. II. Baltimore: Univ. Park Press, 1-14.
- Scherer, K. R. 1974. Beobachtungsverfahren zur Mikroanalyse nonverbaler Verhaltensweisen. In: van Koolwijk, J. & Wieken-Mayser, M. (Hrsg.): *Techniken der empirischen Sozialforschung*. München: Oldenbourg.
- Schmitt, N. 1978. Path analysis of multitrait-multimethod matrices. *Applied Psychological Measurement*, 2, 157-173.
- Schmitt, N., Coyle, B. W. & Saari, B. B. 1977. A review and critique of analyses of multitrait-multimethod matrices. *Multivariate Behavioral Research*, 12, 447-478.
- Schoggen, P. 1976. Ecological psychology and mental retardation. Paper presented at Conference on the Application of Observation/Ethological Methods in the Study of Mental Retardation. Washington, D. C., June.
- Schucany, W. R. & Frawley, W. H. 1973. A rank test for two group concordance. *Psychometrika*, 38, 249-258.
- Schutz, W. C. 1952. Reliability, ambiguity, and content analysis. *Psychological Review*, 59, 119-129.

- Schwartz, M. S. & Schwartz, C. G. 1955. Problems in participant observation. *American Journal of Sociology*, 66.
- Scott, W. A. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19, 321-325.
- Scott, W. A. & Wertheimer, M. 1962. Introduction to psychological research. New York: Wiley, (4. Aufl. 1967).
- Selvage, R. 1976. Comments on the analysis of variance strategy for the computation of intraclass reliability. *Educational and Psychological Measurement*, 36, 605-609.
- Simon, A. & Boyer, E. G. 1974. Mirrors for behavior III. An anthology of Observation instruments. Communication Materials Center, Wyncote, Pennsylvania.
- Simons, G. & Papousek, H. 1978. Methoden der Kleinkindforschung: Beobachtung und Experiment. In: Dollase, R. (Hrsg.): *Handbuch der Früh- und Vorschulpädagogik*. Band 2. Düsseldorf: Schwann, 93-110.
- *Singleton, W. T. 1972. Techniques for determining the causes of error. *Applied Ergonomics*, 3, 126-131.
- *Somers, R. H. 1962. A new asymmetric measure of association. *American Sociological Review*, 1962, 27, 799-811.
- Sorembe, V.: Computer program for and some comments on K. Krippendorff's 'Estimating the reliability, systematic error and random error of interval data'. Institut für Psychologie der RWTH Aachen, Arbeitsbericht, o.J.
- Spearman, C. 1906. „Footrule“ for measuring correlation. *British Journal of Psychology*, 2, 89-108.
- Stanley, J. C. 1961. Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. *Psychometrika*, 26, 205-219.
- Stewart, R. A., Powell, G. E. & Chetwynd, S. J. 1979. Person perception and stereotyping. Westmead: Saxon House.
- Stroud, T. W. F. 1974. Comparing regressions when measurement error variances are known. *Psychometrika*, 39, 53-68.
- Susman, E. J., Peters, D. J. & Steward, R. B. 1976. Naturalistic observational child study, a review. Paper presented at the 4th Biennial Southeastern Conference on Human Development at Nashville, Tennessee.
- Sutcliffe, J. P. 1958. Error of measurement and the sensitivity of a test of significance. *Psychometrika*, 28, 9-17.
- Sutcliffe, J. P. 1980. On the relationship of reliability to statistical power. *Psychological Bulletin*, 88, 509-515.
- Taylor, W. L. 1964. Correcting the average rank correlation coefficient for ties in the rankings. *Journal of the American Statistical Association*, 59, 872-876.
- Taylor, W. L. & Fong, C. 1963. Some contributions to average rank correlation methods and to the distribution of the average rank correlation coefficient. *Journal of the American Statistical Association*, 58, 756-769.

- Tesser, A. & Krauss, H. 1976. On validating a relationship between constructs. *Educational and Psychological Measurement*, 36, 111-121.
- Thomae, H. 1959. Forschungsmethoden der Entwicklungspsychologie. In: Thomae, H. (Hrsg.): *Handbuch der Psychologie*, Band 3, *Entwicklungspsychologie*. Göttingen: Hogrefe, 46-75.
- Thomae, H. 1968. *Das Individuum und seine Welt*. Göttingen: Hogrefe.
- Tinsley, H. E. A. & Weiss, D. J. 1975. Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, 22, 358-376.
- Trippi, R. R. & Settle, R. B. 1976. A nonparametric coefficient of internal consistency. *Multivariate Behavioral Research*, 11, 419-424.
- Tucker, L. R. 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31, 279-311.
- Vegelius, J. 1977. An ordinal scale generalization of the G index invariant over item reflection. *Educational and Psychological Measurement*, 37, 31-35.
- Vegelius, J. 1977. On the weighted G index. *Educational and Psychological Measurement*, (a), 37, 839-842.
- Vegelius, J. 1979. A G index generalization for trichotomized data. *Educational and Psychological Measurement*, 39, 23-27.
- Wackerly, D. D., McClave, J. T. & Rao, P. V. 1978. Measuring nominal scale agreement between a judge and a known standard. *Psychometrika*, 43, 213-223.
- Watkins, M. W. & McDermott, P. A. 1979. A Computer program for measuring levels of Overall and partial congruence among multiple observers on nominal scales. *Educational and Psychological Measurement*, 39, 235-239.
- *Webb, E. J., Campbell, D. T., Schwartz, R. D. & Sechrest, L. 1966. *Unobtrusive measures: nonreactive research in the social sciences*. Chicago: Rand McNally.
- Weick, K. E. 1968. Systematic observational methods. In: Lindzey, G. & Aronson, E. (eds): *The handbook of social psychology*. Vol. 2. Reading, Mass.: Addison-Wesley, 2. Aufl., 357-451.
- Werner, J. 1976. Varianzanalytische Maße zur Reliabilitätsbestimmung von ratings. *Zeitschrift für experimentelle und angewandte Psychologie*, 23, 489-500.
- Werts, C. L. & Linn, R. L. 1970. Path analysis: Psychological examples. *Psychological Bulletin*, 74, 193-212.
- Werts, C. E. & Linn, R. L. 1971. Analyzing school effects: Ancova with a fallible covariate. *Educational and Psychological Measurement*, 31, 95-105.
- Werts, C. E., Jöreskog, K. G. & Linn, R. L. 1971. Comment on 'The Estimation of Measurement Error in Panel Data'. *American Sociological Review*, 36, 110-113.
- *Werts, C. E., Jöreskog, K. G. & Linn, R. L. 1973. Identification and estimation in path analysis with unmeasured variables. *American Journal of Sociology*, 78, 1469-1484.

- Werts, C. E., Linn, R. L. & Jöreskog, K. G. 1974. Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement*, 34, 25-33.
- Wheaton, B., Mutherr, B., Alwin, D. F. & Summers, G. F. 1977. Assessing reliability and stability in panel models. In: Heise, D. R. (ed.): *Sociological Methodology* 1977. San Francisco: Jossey-Bass, 84-136.
- *Whelan, P. 1974. Reliability of human observers. Doctoral Diss., Univ. of Utah at Salt Lake.
- Whiting, J. W. M., Child, I. L. & Lambert, W. W. 1966. Field guide for a study of socialization: six cultures series, Vol. I. New York: Wiley.
- Whyte, W. F. 1953. Observational field work methods. In: Jahoda, M., Deutsch, M. & Cook, S. W. (eds): *Research methods in the social sciences*. New York: Dryden, 493-513.
- Wiley, D. E. & Wiley, J. A. 1970. The estimation of measurement error in panel data. *American Sociological Review*, 35, 112-117. (Nachdruck in: Blalock, H. M. (ed.): *Causal models in the social sciences*. London: Macmillan, 1971).
- Wilson, T. P. 1974. Measures of association for bivariate ordinal hypothesis. In: Blalock, H. M. (ed.): *Measurement in the social sciences*. Chicago: Aldine-Atherton, 327-342.
- Winer, B. J.: *Statistical principles in experimental design*. New York: McGraw-Hill, 1. Ausg. 1962; 2. Ausg. 1971.
- Wright, H. F. 1960. Observational child study. In: Mussen, P. (ed.): *Handbook of research methods in child development*. New York: Wiley, 71-139.
- Yates, F. 1953. *Sampling methods for censuses and surveys*. London: Griffirr, 2. Aufl.

2. Kapitel

Beobachtung und Beschreibung von Erleben und Verhalten

Hubert Feger und Carl F. Graumann

1. Vorbemerkungen zu Thema und Terminologie

Wenn Erleben und Verhalten die zentralen Begriffe sind, die den Gegenstand der Psychologie umschreiben, und wenn vor jeder weiteren Datenanalyse Beobachtungen anzustellen und diese zu beschreiben sind, darf man dann nicht eine intensive und traditionsreiche Auseinandersetzung über die Beziehungen zwischen Beobachten, Beschreiben, Erleben und Verhalten erwarten? Bekanntlich ist der Diskussionsstand sehr unausgeglich, und wir können dort keine Systematik berichten, wo der Forschungsstand zu viele unverbundene Probleme nur nebeneinander stellt.

Wir gehen davon aus, daß sowohl Erleben als auch Verhalten sowohl beobachtet als auch beschrieben werden können. Wenn es derjenige selbst ist, der erlebt und sich verhält, der dies beobachtet und beschreibt, so können wir von Selbstberichten und Selbstbeschreibungen sprechen (self report, self recording). Geschieht Beobachtung und Beschreibung durch einen oder mehrere andere, so können wir das als Fremdbeobachtung bezeichnen. Die erste systematische Schwierigkeit beginnt mit dem Einwand, ihr Erleben könne eine Person nur selbst beobachten. Demnach gehört zum Thema dieses Kapitels:

1. *Die Beobachtung des eigenen Erlebens*, meist, synonym, als Selbstbeobachtung und Erlebnisbeschreibung bezeichnet (wobei unklar bleibt, warum in der Literatur nicht auch durchgängig und synonym von Erlebnisbeobachtung und Selbstbeschreibung die Rede ist). Erkenntnis- und wissenschaftstheoretische Arbeiten sind in diesem Bereich relativ zahlreich; nur einige können wir streifen. Methodische Arbeiten sind hingegen ausgesprochen selten, sie finden sich nicht einmal gehäuft zur Blütezeit des „Introspektionismus“. Den Terminus Introspektion wollen wir nur dann für Selbstbeobachtung gebrauchen, wenn zugleich entsprechende bewußtseinstheoretische Positionen (s.u.) mitgemeint sind. Andererseits hat auch Skinners Hypothese, die Möglichkeit „private

events“ zu beschreiben, beruhe auf sozialem Lernen, ebenfalls nicht zu einer Serie gezielter Experimente geführt, die diesen Prozeß untersucht und die Verzahnung innerer Zustände und angebotenen Beschreibungsarsenal genauer analysiert hätten.

2. *Die Beobachtung des eigenen Verhaltens* (wobei strittig ist, wie berechtigt es bei Selbstberichten ist, zwischen Erleben und Verhalten scharf zu trennen).

3. *Die Beobachtung fremden Verhaltens* (gegenwärtig im therapeutischen Kontext oft als Verhaltens einschätzung akzentuiert und bezeichnet). Für die wissenschaftliche Gemeinschaft liegt stets nur das Produkt, nicht der Prozeß der Beobachtung in Form von Beschreibungen i. w. S. des Wortes vor. Empirische psychologische Arbeiten über den Zusammenhang zwischen Beschreiben und Beobachten sind selten; methodische Arbeiten über Verhaltensbeobachtung, besonders über eher technische Fragen, aber auch über die Psychologie der Verhaltensbeobachtung, sind in jüngerer Zeit häufiger. Das folgende Kapitel kann nicht mehr als eine Auswahl der in der Literatur bearbeiteten Thementeile ansprechen.

2. Formen der Erlebnisbeschreibung

2.1 Selbstbeobachtung und Erlebnisbeschreibung als Methoden und Themen der Psychologie

Die Selbstbeobachtung ist dem Alltagsverstand bekannt, aber nicht immer ganz geheuer, noch ist es die Beobachtung, die einer über einen anderen anstellt.

Unter dem Titel „Von dem Beobachten seiner selbst“ schreibt ein einflußreicher und für das Schicksal der wissenschaftlichen Psychologie maßgeblicher Denker zur Zeit der vorletzten Jahrhundertwende:

„Das Bemerken (animadvertere) ist noch nicht ein *Beobachten* (observare) seiner selbst. Das letztere ist eine methodische Zusammenstellung der an uns selbst gemachten Wahrnehmungen, welche den Stoff zum *Tagebuch eines Beobachtens seiner selbst* abgibt und leichthin zu Schwärmerei und Wahnsinn einführt.“

Der Autor und die Quelle, Immanuel Kant und seine „Anthropologie“ von 1800 (1. Aufl. 1798), schließen bei aller humanistischen Diktion und pathologischen Verweisung wohl jede Assoziation mit zeitgenössischer Wiederverwendung der Selbst-Beobachtung und des Tagebuchs als Mittel einer „humanistisch“ orientierten klinischen Psychologie zweifelsfrei aus. Wohl aber belegt das wohl unmißverständliche Zitat die Tatsache, daß der Selbstbeobach-

tung vor aller psychologisch-methodologischen Spezifizierung eine anthropologische Allgemeinheit zuzusprechen ist.

Man darf dabei aber nicht übersehen, daß, von der Beobachtung seiner selbst zu reden, nicht unbedingt und nicht einmal in erster Linie identisch ist mit dem, was anderswo und seit geraumer Zeit auch bei uns mit „*Introspektion*“ bezeichnet wird, von wo ausgehend der „Introspektionismus“ die (theoretische?) Richtung ist, von der kein Vertreter je existierte, der sich selbst so verstanden und benannt hätte, Introspektion, wörtlich „Innenschau“, meint, unter der doppelten Voraussetzung einer cartesischen Dichotomie von Innen- und Außenwelt als „kognitiver“ und „ausgedehnter“ Substanz und eines analog (oder parallel?) zum äußeren Auge gedachten inneren Auges, der Blick nach innen auf das eigene Bewußtsein, sofern - aber auch nur sofern - *Bewußtsein als rein Inneres* auffaßbar erscheint (vgl. hierzu Graumann 1966). Fassen wir hingegen Bewußtsein im phänomenologischen Sinne intentional, d.h. immer als Bewußtsein oder Erleben von etwas Inner- oder Außerweltlichem, dann ist alles „da draußen“ Wahrgenommene, Vorgestellte, Vermutete, Erinnernte etc. zwar als Erlebtes auch prinzipiell erlebnisdeskriptiv konstatierbar, aber nicht „introspezierbar“. Die vor allem in der behavioristischen Literatur der USA anzutreffende Gleichsetzung von „phänomenal“ und „introspektiv“ erscheint angesichts etwa der Wahrnehmung „blauer Berge“ und ähnlicher Erlebnisse unvertretbar.

Nun würde jemand, der die „blauen Berge“ aus der Ferne erlebt hat, nicht nur überrascht sein, wenn man ihm dieses Erlebnis als „Introspektion“ anrechnete; er würde ebenso lebhaft protestieren, wenn ihm dieser wundervolle Blick in die Weite als „Selbstbeobachtung“ attestiert würde. Vergißt man für einen Augenblick die - wie man sieht - nicht immer klärende Terminologie, dann beobachten wir uns doch am ehesten selbst, wenn wir anfangen, über uns zu stutzen und nachzudenken. Das kann durchaus - in einem ursprünglichen Sinne des Wortes Selbstbeobachtung - angesichts des allmorgendlichen Spiegelbildes passieren. Und der Verdacht, daß, relativ und zwar „negativ“ zum gewohnten Bild, heute Auge, Lippen oder das Ganze anders aussehe, kann zur Steigerung der Aufmerksamkeit beim Hinsehen führen, ohne daß nach den Spielregeln der Psychoanalyse bereits „Narzißmus“ zu diagnostizieren wäre. Auch dann würden wir ohne Bedenken von einer Selbstbeobachtung reden, wenn einer, der kritisiert wurde, weil er ständig „nich' wahr“ sagt, beim Reden darauf achtet und dann, solange er darauf achtet, es auch nicht hört, weil nicht mehr sagt. Denn Beobachten heißt ja nur so viel wie ein intensives Achten-auf, und Auf-sich-selbst-Achten tut man, wenn dazu Anlaß oder Grund bestehen und einem eine entsprechende Paß-auf-Instruktion von einem selbst oder anderen erteilt worden ist. In diesem allgemein verständlichen und klaren Sinn unterscheiden sich alltägliche und wissenschaftliche Selbstbeobachtung nicht; nicht einmal die Zweifel an der Tauglichkeit dieses Vorgehens! Denn, ob ich

das, was an meinem Verhalten zu beobachten ist, auch tatsächlich beobachte, bleibt immer offen, und ich erfahre nie verlässlich, ob es an der Beobachtungsabsicht oder am Ausbleiben des fraglichen Phänomens liegt, wenn ich das Erwartete nicht beobachte.

Jenseits dieser auch der Alltagserfahrung vertrauten Selbstbeobachtung, die, habitualisiert, wie schon Kant wußte, auch „krankhaft“ werden kann, liegt das Grübeln, die Reflexion über sich selbst. Warum habe ich nur dies getan, das gelassen? Was in aller Welt hat mich dazu gebracht, getrieben etc.? - Fragen, die man sich selbst über die Ursachen oder Gründe des eigenen (aber genauso: fremden) Tuns stellt. Diese Form der Retrospektion sollte als eine Form des Nachdenkens über sich selbst (und die Welt) nicht mit dem gleichen Begriff gefaßt werden wie die mehr oder minder unmittelbare Beobachtung eigenen Seins oder Tuns.

Wenn heute gerne unscharf vom „verbalen Report“ (statt von Selbstbeobachtung oder Erlebnisbeschreibung) die Rede ist, geht dieser Unterschied zwischen dem unmittelbaren Achten-auf und dem u.U. räsonierenden Nachdenken-über (mutmaßliche) Zusammenhänge verloren; es zählt nur das „verbale“ Resultat. Ein „verbaler Report“ als Antwort auf Warum-Fragen (vgl. Nisbett & Wilson 1977) kann sehr unterschiedliche Quellen haben; eine davon mag das eigene Erleben sein, beziehungsweise die Art und Weise, wie wir über unsere Erlebnisse, über unser Inneres, unsere Innerlichkeit, über das ganz Private etc. zu reden gelernt haben, sei es gegenüber nahestehenden Vertrauten, gegenüber Fremden oder auch gegenüber sich „persönlich“ gebenden Fremden wie Psychologen.

Damit bleibt als letztes und sicher nicht geringstes Problem die von Psychologen gerne und bewußt vollzogene Gleichsetzung von Beobachtung und Beschreibung, hier also von „Selbstbeobachtung“ und „Erlebnisbeschreibung“. Zwar lehrt uns die Anwendung der Datentheorie auf die Beobachtungsmethodik, daß eine Beobachtung, was ihre wissenschaftliche Güte betrifft, nie besser sein kann als ihre Kategorisierung. Aber dieser stolze Satz klingt weit weniger überzeugend, wenn wir kein verlässliches Verfahren aufweisen können, daß die Beziehung zwischen dem Phänomen und seiner Kategorisierung zu präzisieren gestattet. Falls es diese Beziehung überhaupt gibt zwischen „Anschauung“ und „Begriff“, wird es so viele Methodenklassen geben, wie Modalitäten dieser Beziehung denkbar sind. Die Flucht vor diesem Problem, das die Geschichte der abendländischen Philosophie mitgeprägt hat, besteht im Rückzug auf die Rede, auf den Diskurs: Bescheiden wir uns mit der Art und Weise, wie Leute über bestimmte Themen reden und analysieren diese Rede ohne Rekurs, sei es auf „Tatsachen“, sei es auf „Erlebnisse“.

Wenn dies einige unterscheidbare Modi der kognitiven Beschäftigung des Individuums mit sich selbst sind, dann sollten sie auch Themen einer Wissenschaft

darstellen, die das Erleben und Verhalten von Menschen zum Gegenstand hat. Tatsächlich reichen seit einiger Zeit, nimmt man nur Persönlichkeits- und Sozialpsychologie, umfangreiche Forschungsthemen von der Selbstaufmerksamkeit und Selbstwahrnehmung über die Selbstattribution und Selbstbeurteilung bis hin zur Selbstdarstellung im Alltagsleben. Sie alle verdanken ihre Existenz dem anthropologischen Sachverhalt der *Reflexivität des menschlichen Bewußtseins*. Das heißt, daß der einzelne nicht nur seine Umwelt und Mitwelt, sondern auch sich selbst in den verschiedenen Modalitäten der Wahrnehmung, Erinnerung, Antizipation, des Empfindens und Fühlens, des Denkens und Urteilens erfährt (erlebt) und anderen darüber direkt oder indirekt Aussagen machen kann.

Diese, sagen wir, doppelte Fähigkeit, sich selbst in verschiedenen Modalitäten zu erleben und darüber hinreichend differenzierende Aussagen zu machen, ergibt nicht nur eine ganze Klasse von problemgeladenen Forschungsthemen. Sie konstituiert auch die Klasse(n) von wissenschaftlichen Verfahren, die Aussagen des Menschen über sich selbst zum Ausgang nehmen. Diese Verfahren können sehr unterschiedlichen (wie diagnostischen, klinischen) Zwecken dienen. In diesem Abschnitt werden sie nur insofern behandelt, als sie *erlebnisdeskriptiv* sein bzw. Erlebnisdeskription ermöglichen sollen.

Dem Hinweis darauf, daß der kognitiv-sprachliche Rückzug auf sich selbst bzw. auf das eigene Erleben, der seit Beginn der wissenschaftlichen Psychologie als Forschungsmethode diente, inzwischen auch als Forschungsthema volle Aufmerksamkeit findet, kann man bereits entnehmen, daß am Kernproblem der Erlebnisdeskription, nämlich der Beziehung von Erlebnis und Deskription, noch gearbeitet wird.

2.2 Selbstbeobachtung und Experiment:

Die Begründung der wissenschaftlichen Psychologie

Blickt man auf die Anfänge der wissenschaftlichen Psychologie im neunzehnten Jahrhundert zurück, so wird erkennbar, daß die Begründung der Psychologie als Einzelwissenschaft eine methodologische war. Zumindest für Wilhelm Wundt gilt, daß er die Überzeugung, „daß die Fortschritte jeder Wissenschaft innig an die Fortschritte der Untersuchungsmethoden gebunden sind“ (Wundt 1862, xi), in einer für lange Zeit verbindlichen Weise in die Tat umgesetzt hat. Worin aber sah Wundt den für die Psychologie so dringend nötigen Fortschritt? Was immer schon, vorwissenschaftlich - und das heißt vor allem: philosophisch - Psychologie treiben ausmachte, war die Selbstwahrnehmung oder Selbstbeobachtung. Generationen von Philosophen, aber auch Seelenkundlern, hatten aus der unmittelbaren Erfahrung ihrer eigenen Empfindungen und Ideen, Affekte und Begierden und der Reflektion darüber ihre Psy-

chologie geschöpft. Nur war darüber Psychologie, zumindest im Vergleich zu der sich kräftig entwickelnden Naturwissenschaft, nicht zur Wissenschaft geworden. Gleichwohl, als Wundt dies im Sinne der Begründung einer Psychologie als Naturwissenschaft zu ändern sich anschickte, blieb auch für ihn gültig: „Alle Psychologie beginnt mit der Selbstbeobachtung, und diese bleibt zur Beurteilung der außer uns stehenden psychischen Erscheinungen immer ein unentbehrliches Hilfsmittel“ (a. a. O., xvi). Allerdings, so fügt Wundt hinzu, ist sie völlig unzureichend für die Entwicklungs- und Kausalanalyse der Erscheinungen.

Daß wir der Selbstbeobachtung geradezu auch bedürfen, um „außer uns stehende“ psychische Erscheinungen überhaupt verstehen zu können, vertrat auch - hier, wie so oft, wundtscher als Wundt - noch sechzig Jahre später E.B. Titchener (1914, 32), wenn er behauptet, daß wir „immer wieder auf die experimentelle Selbstbeobachtung zurückgreifen müssen“, wenn wir den Versuch machten, „die psychischen Prozesse eines Kindes oder eines Hundes oder eines Insekts“ zu verstehen, wie sie sich an äußeren Verhaltensmerkmalen zu erkennen geben. „Wir können uns keine Prozesse in einem anderen Bewußtsein vorstellen, die wir nicht in unserem eigenen finden.“ Fazit: „Experimental introspection . . . is the sole gateway to psychology.“

Auch William James, um einen der Wundt-Anhängerschaft unverdächtigen Mitbegründer der modernen Psychologie zu zitieren, beginnt sein Kapitel über die Forschungsmethoden der (ausdrücklich:) Naturwissenschaft Psychologie mit dem hervorgehobenen Satz: *Introspective observation is what we have to rely on first and foremost and always*“ und fügt hinzu, das Wort Introspektion bedürfe wohl keiner Definition, es meine selbstverständlich „looking into our own minds and reporting what we there discover. Every one agrees that we there discover states of consciousness“ (James 1890, I, 185).

Daß wir Bewußtsein irgendwelcher Art haben, nennt James das *inconcussum* in einer Welt, wo alles andere sich als bezweifelbar erwiesen hat. Entsprechend wird die Überzeugung, daß wir Bewußtsein haben, und auch, daß wir unser Bewußtsein von dem unterscheiden können, was Gegenstand dieses Bewußtseins werden kann, als „das grundlegendste aller Postulate der Psychologie“ bezeichnet (ebda.).

Wir nehmen zwar die Begründer der modernen Psychologie als Ausgangsbeispiele für die enge Verknüpfung von Psychologie und Selbstbeobachtung, legen aber Wert darauf, daß diese Konzeption, vor allem die von naturwissenschaftlicher Psychologie und Selbstbeobachtungsmethode, keine historische Episode geblieben ist. Bevor wir uns den Formen und Problemen der Selbstbeobachtungsmethode bzw. der Erlebnisdiskription zuwenden, die auch die Gegenwart beschäftigen, sei noch einer der einflußreichsten deutschsprachigen Psychologen der Nachkriegsjahre bemüht, der ebenfalls Psychologie als Na-

turwissenschaft vertrat, Hubert Rohrer. Für ihn läßt sich gleichwohl ein grundsätzlicher Unterschied zwischen den Forschungsmethoden der Psychologie und der übrigen Wissenschaften formulieren: „In allen anderen Wissenschaften hat man Dinge oder Vorgänge zu untersuchen, die außerhalb des eigenen Erlebens liegen, während man in der Psychologie gezwungen ist, die eigenen Erlebnisse zu beschreiben; das Objekt der Psychologie - das bewußte Erleben - findet man nirgendwo anders als in sich selbst.. *Die wichtigste Methode der Psychologie ist daher die Selbstbeobachtung*“¹⁾ (Rohrer 1963, 69 f.).

Allen, denen es um die Begründung oder Bewahrung der Psychologie als Wissenschaft, speziell als Naturwissenschaft zu tun war, ist die Problematik der Selbstbeobachtung vertraut gewesen. Sie hatte bereits manchen Denker, so vor allem Kant (1786; 1800) und Comte (1830-1842) dazu gebracht, Psychologie als Wissenschaft für unmöglich zu halten - immer unterstellt, daß sich eine den Namen Psychologie tragende Wissenschaft mit dem Psychischen (oder dem Bewußtsein oder dem Erleben) notwendig und letztlich zu beschäftigen habe.

Insofern sollte es nicht überraschen, bei den Befürwortern der Selbstbeobachtung als grundlegender Methode der Psychologie auch die kritischsten Aussagen über diese Methode zu finden. So hat es Sekundärkenner von Wundt immer wieder überrascht, neben dem im obigen Zitat zum Ausdruck kommenden Bekenntnis zur Selbstbeobachtung bei Wundt auch Sätze zu finden, wonach „Selbstbeobachtung, wenn wir das Wort Beobachtung im wissenschaftlichen Sinne verstehen, unmöglich ist. Je mehr wir uns anstrengen, uns selber zu beobachten, um so sicherer können wir sein, daß wir überhaupt gar nicht beobachten“ (Wundt, 1906, 196).

Was hier widersprüchlich erscheint, erklärt sich nicht daraus, daß sich Wundts Methodik-Konzept von seiner Heidelberger zu seiner Leipziger Zeit geändert hat. Das ist zwar auch der Fall (Graumann 1980), betrifft aber nicht die Selbstbeobachtung. Die scheinbar widersprüchlichen Aussagen sind, wie Blumenthal (1975) und Metge (1980) erneut gegen tradierte Mißverständnisse der Wundtschen Methodik demonstrieren, voll vereinbar. Man muß sich nur deutlich machen, daß Wundt von Anfang an die reine Selbstbeobachtung für methodisch wertlos, die mit dem experimentellen Verfahren gepaarte Selbstbeobachtung (besser allerdings: Selbstwahrnehmung) jedoch für unabdingbar ansah. Es war die nicht (hinreichend) kontrollierbare und damit letztlich nicht verifizierbare Selbstbeobachtung, die er anfänglich den vorwissenschaftlichen Psychologen (Wundt 1862; 1888), später den Denkpsychologen der Kälpe-Schule (S.U.) zum Vorwurf machte (Wundt 1907).

¹⁾ Bei Rohrer ist nur das letzte Wort hervorgehoben.

Um die für seine Psychologie unverzichtbare Selbstwahrnehmung in eine wissenschaftliche Methode zu transformieren, muß sie der Kontrolle des Experiments unterworfen werden, wobei unter Experiment, in der Tradition von Francis Bacon (1620; 1974), ganz zu Anfang und - wir meinen - doch in verbindlicher Weise ein dreifach kontrollierendes Verfahren verstanden wird (Wundt 1863, I, Vf.):

„Durch das Experiment erzeugen wir die Erscheinung künstlich aus den Bedingungen heraus, die wir in der Hand halten. Wir verändern diese Bedingungen und verändern dadurch in meßbarer Weise auch die Erscheinung.“

Da die Veränderung von Bedingungen die Wiederholung konstanter Bedingungen (und sei es nur für Kontrollzwecke), wie Wundt später selbst sah, zur Voraussetzung hat, läßt sich konsequenterweise von einem (durch Manipulierbarkeit, Wiederholbarkeit, Variierbarkeit und Meßbarkeit) vierfach kontrollierenden Verfahren reden. Allerdings hat Wundt, für den das Experiment eine Form der Beobachtung war, später das strengere Kriterium der Meßbarkeit durch das der (aufmerksamen) Beobachtbarkeit ersetzt (vgl. hierzu Graumann 1980, 76f.). Für die methodische Selbstbeobachtung ergab sich dann der folgende Regelkanon:

- (1) „Der Beobachter muß, wo möglich, in der Lage sein, den Eintritt des zu beobachtenden Vorgangs selbst bestimmen zu können.
- (2) Der Beobachter muß, soweit möglich, im Zustand gespannter Aufmerksamkeit die Erscheinungen auffassen und in ihrem Verlauf verfolgen.
- (3) Jede Beobachtung muß zum Zweck der Sicherung der Ergebnisse unter den gleichen Umständen mehrmals wiederholt werden können.
- (4) Die Bedingungen, unter denen die Erscheinung eintritt, müssen durch Variation der begleitenden Umstände ermittelt und . . . in den verschiedenen zusammengehörigen Versuchen planmäßig verändert werden, indem man sie teils in einzelnen Versuchen ganz ausschaltet, teils in ihrer Stärke oder Qualität abstuft“ (Wundt 1907, 308).

Erscheinen hier die formalen Kriterien des Experiments in der Funktion, wissenschaftliche von unwissenschaftlicher (Selbst-) Beobachtung zu scheiden, so gilt, was theoretisch bedeutsam ist, umgekehrt der Einsatz der Selbstbeobachtung als Legitimation des Experiments in der Psychologie. Vergegenwärtigt man sich die Fragestellungen der „Physiologischen Psychologie“, wurde tatsächlich jedes in diesem Kontext entworfene Experiment, das auf Selbstbeobachtung verzichtete, zum physiologischen Experiment (vgl. Metge, a.a.O., 186).

Es dürfte aus den vier zitierten Regeln, die im einzelnen als Kann-Bestimmungen galten, erkennbar sein, daß die Einsatzmöglichkeiten wissenschaftlicher Selbstbeobachtung begrenzt waren auf relativ einfache, d.h. „überschaubare“,

Vorgänge, die zum Zwecke der objektiven Kontrolle möglichst auf physische Reize bezogen sein sollten, auch wenn reine Erlebnisbeobachtung zulässig war, sofern einige der Regeln eingehalten wurden. Das aber heißt, daß schon die Selbstbeobachtungsmethodik der meisten Wundt-Schüler als nicht mehr regelgerecht galt.

2.3 Die systematische experimentelle Selbstbeobachtung

2.3.1 Die konkrete Vorgehensweise

Im Gegensatz zur heute ausgefeilten Methodik etwa der Testkonstruktion oder der Versuchsplanung gibt es keine „Einführung in die Methodik der Selbstbeobachtung“. Vielmehr scheint es so, als ob die sie anwendenden Psychologen und ihre Versuchspartner durchweg glauben, man beherrsche Selbstbeobachtung natürlicherweise, als ob es einer Methodenlehre der Selbstbeobachtung nicht bedürfe, allenfalls einer Grenzbestimmung ihrer Möglichkeiten. Jedoch erfordert auch eine solche abgrenzende Kritik eine systematische methodische Auseinandersetzung mit dieser empirischen Vorgehensweise, und so haben es auch die Psychologen um die Jahrhundertwende gesehen. Aus ihrem konkreten Vorgehen läßt sich ihre implizite Methodenlehre entwickeln und dann diskutieren. Wir schildern daher als erstes ihr methodisches Vorgehen, wobei wir eine der sorgfältigsten Anwendungen zugrunde legen, die durch N. Ach (1905).

Ach (S. 8) bestimmt als *Ziel* der experimentellen Selbstbeobachtung, „ein vollständiges, zuverlässiges und unbefangenes Bild der wirklich vorhandenen Bewußtseinsinhalte“ zu geben. Dieses Ziel erreicht die „systematische experimentelle Selbstbeobachtung“, allgemein beschrieben, auf folgendem Weg (Ach, 1905, S.8f.):

„Die Methode der systematischen experimentellen Selbstbeobachtung geht, wie bereits bemerkt, darauf aus, das durch äußere experimentelle Hilfsmittel veranlaßte Erlebnis der Versuchsperson jedesmal in der dem Versuche unmittelbar folgenden Zeit einer vollständigen Beschreibung und Analyse zu unterwerfen. Hierbei findet ein fortwährender enger Gedankenaustausch zwischen der beobachtenden Versuchsperson und dem protokollierenden Versuchsleiter statt. Da jede Versuchsanordnung im allgemeinen durch ein vorbereitendes Signal, welches die notwendige Einstellung der Aufmerksamkeit bezweckt, eingeleitet wird, so lassen sich beim psychologischen Einzelversuch drei Zeitabschnitte unterscheiden:

- 1) die *Vorperiode*, welche die Zeit zwischen Signal und Eintritt des Reizes umfaßt,
- 2) die *Hauptperiode*, welche das eigentliche experimentell zu untersuchende Erlebnis in sich schließt,
- 3) die *Nachperiode*, welche die sich unmittelbar an den Abschluß des Experimentes anschließende Zeit umfaßt.

Der gesamten jeweiligen Versuchsreihe hat außerdem die Angabe der Instruktion vorauszugehen. Die Instruktion der Versuchsperson hinsichtlich der Selbstbeobachtung lautet dahin, die in der Vorperiode und Hauptperiode erlebten Vorgänge in der Nachperiode eingehend zu schildern. Selbstverständlich hatte die Versuchsperson auch die Pflicht, bemerkenswerte Erlebnisse in den Zwischenpausen zwischen den einzelnen Versuchen, so eine stattfindende associative Einübung u. dergl. dem Versuchsleiter anzugeben.“

Es wird deutlich, daß Ach hier Selbstbeobachtung als Aufgabe, zeitlich unmittelbar vorausgegangenes Erleben zu beschreiben, auffaßt, so wie Stern (1911, S. 38) als Aufgabe der Selbstbeobachtung vorgibt „... die Feststellung eines *akuten Merkmals* (Hervorhebung von Stern), d.h. eines bestimmten, in einem gegebenen Augenblick vorhandenen Phänomens oder Aktes in der sich beobachtenden Persönlichkeit.“ Davon unterschieden wird Selbstbeurteilung, die als Aufgabe des Forschers nach Bedingungen für psychische Prozesse und Strukturen sucht.

2.3.2 Maßnahmen zur Sicherung der Ergebnisse

Der Versuchsleiter hat im Rahmen dieses allgemeinen Schemas bestimmte Maßnahmen zu ergreifen, die gewährleisten sollen, daß die Vp ihre Erlebnisse vollständig, zuverlässig und unverfälscht schildert. Dazu gehört als erstes, die Nachperiode von den beiden anderen zu trennen, und nur in der Nachperiode das Erlebte zu schildern, und diese Sequenz öfters zu wiederholen, damit die *Beobachtung des Erlebens oder schon die Absicht zu beobachten den Erlebnisablauf oder seinen Inhalt* nicht stören kann. Da dieses Argument bis heute zu den zentralen in der Auseinandersetzung um die Selbstbeobachtung gehört, geben wir es vollständig wieder (Ach 1905, S.9):

„Es fällt hiermit jener Einwand weg, der schon von Kant (Ach verweist auf: *Metaphysische Anfangsgründe der Naturwissenschaft*. Vorrede S. XI, 1786) angedeutet und seitdem häufig wiederholt wurde, daß eine direkte Beobachtung der psychologischen Phänomene während ihres Erlebtwerdens oder die Absicht, während des Vorganges zu beobachten, den zu untersuchenden Prozeß unmöglich macht. Denn hier findet während des Erlebnisses für gewöhnlich keine Beobachtung statt, ebensowenig besteht die Absicht, während des Erlebens das zu untersuchende Geschehen zu beobachten. Daß die Selbstbeobachtung auf das Erlebnis, solange dasselbe sich nicht öfters wiederholt hat, einen störenden Einfluß ausübt, davon konnte ich mich bei meinen Untersuchungen vielfach überzeugen.“

Um eine *verfälschende Auswahl* durch die Vp aus ihrem Erleben zu vermeiden und um die Vp in *Unwissenheit* darüber zu lassen, woran genau der VI interessiert ist, fordert Ach, jedesmal das gesamte Erlebnis so vollständig wie nur möglich zu beobachten und zu berichten, insbesondere nicht nur das gerade wichtig erscheinende oder lebhaft hervortretende (S. 14). Da die gewünschte Vollständigkeit „bei den überaus reichhaltigen psychischen Erlebnissen“ sich

von keiner Vp erreichen lasse, seien Wiederholungen nötig, bei denen man nach und nach der Vollständigkeit näher komme (S. 16f.).

Drittens fordert Ach eine Kontrolle darüber, ob die von der Vp gewählte „sprachliche Bezeichnung wirklich den adäquaten Ausdruck des zugehörigen geistigen Inhaltes darstellt“. Ach sieht das Problem in erster Linie darin, daß der *Sprachgebrauch* von Vp und VI nicht übereinstimmen, weniger darin, daß die Sprache grundsätzliche Mängel bei der Abbildung von Erlebten aufweisen könnte (s. Linschoten 1959, 1961). Die fehlende Übereinstimmung läßt sich durch *Nachfragen* klären und vielleicht beseitigen: „Der Versuchsleiter hat deshalb die Pflicht, die gegebene Schilderung durch Fragestellungen zu ergänzen.“ (S. 14). Als Beispiele finden wir:

„Die Fragestellungen bezogen sich auf die zeitliche Aufeinanderfolge, so daß z.B. nach der Schilderung der Versuchsperson gefragt wurde: was ging diesem Zustande vorher? was war zwischen diesen beiden Vorgängen? schlossen sie sich unmittelbar aneinander? standen sie in irgendeiner bewußten Beziehung? Auch der simultane Inhalt wurde in ähnlicher Weise besprochen, z.B. waren die Vorgänge gleichzeitig im Bewußtsein? welchem war die Aufmerksamkeit zugewendet? wie war der Vorgang im Bewußtsein? was für Merkmale hat dieser Vorgang? waren Gefühle dabei u.s.w.? Ist der Vorgang gleich einem vorhergehenden Vorgang? .“ (S. 17).

Viertens, gewissermaßen als zusätzliche Sicherung gegen „Täuschung“, verwendet Ach fast nur solche Beobachtungen, „... welche bei verschiedenen Versuchspersonen übereinstimmend gefunden wurden“ (S.20), und die *inter-individuelle Replikation* geschieht nicht über beliebige Vpn, sondern möglichst an *Psychologen*, wegen der „Schwierigkeit der Durchführung“, und auch bei den Psychologen hat „Schulung und stetige Kontrolle“ stattzufinden (S.23). Das Thema „Übung in der Selbstbeobachtung“ ist bis heute nicht systematisch empirisch untersucht worden.

Unabdingbar ist fünftens, daß es sich um Selbstbeobachtung unter *experimentell variierten Bedingungen* handelt, und:

„Die systematische experimentelle Selbstbeobachtung hat jedoch keinen Wert, wenn es nicht gelingt, durch Änderung der äußeren Versuchsanordnung und der Instruktion auch eine dem jeweiligen Zwecke entsprechende Änderung des inneren Erlebnisses herbeizuführen, und so durch Variierung der äußeren Umstände auch eine Kontrolle der in der Selbstbeobachtung gemachten Angaben durchzuführen.“

Der Forscher muß also Annahmen darüber formulieren, wie äußere Bedingungen und Erleben zusammenhängen, und ihre Bestätigung stärkt das Vertrauen in die verwendete Methode. Der konsequent weitergehende Schluß, ausbleibende Bestätigungen könnten nicht nur zu einer Revision der Theorie, sondern auch der Methodik führen, wurde im Detail in der vorbehavioristischen Zeit nicht gründlich genug, und mit dem Behaviorismus allzu vehement gezogen.

Es gibt, bei Ach selbst wie bei seinen Zeitgenossen, zahlreiche Varianten der Methode, z.B. mit und ohne Fragen des VI, mit und ohne Vorphase, häufig ohne experimentelle Variation der situativen Bedingungen usw.; diese Varianten müssen jedoch nicht diskutiert werden, wenn die grundsätzliche methodische Problematik aufgezeigt werden soll.

2.3.3 Begründung der Möglichkeit von Selbstbeobachtung

Wie jede andere Methode der Datenerhebung bedarf auch die der Selbstbeobachtung des Nachweises, daß die Beobachtungen, die man mit ihrer Hilfe gewinnen will, prinzipiell auch gewonnen werden können. Das Kriterium für einen solchen Nachweis besteht darin zu zeigen, daß (empirisch gesicherte) theoretische Annahmen über den zu erfassenden Gegenstandsbereich, hier psychische Strukturen und Prozesse des Erlebens, und implizite Voraussetzungen der Methode über die Natur dieses Gegenstandsbereiches miteinander vereinbar sind.

Zu den notwendigen psychologischen Annahmen gehört: Die experimentell erzeugte Situation führt an der gleichen Person - von gesetzmäßigen oder kontrollierbaren Reihenfolgewirkungen einmal abgesehen - zu einem Erlebnis, das auch über verschiedene Replikationen soweit identisch ist, daß sich seine Behandlung als „gleiches Phänomen“ theoretisch rechtfertigen läßt. Für die Untersuchung von Erlebnissen - etwa im Gegensatz zu offen beobachtbaren Verhaltensweisen - ist diese Forderung deshalb kritisch, weil über sein Erleben nur der eine Beobachter Auskunft geben kann. Ferner muß man annehmen (Ach, S. 10), Selbstbeobachtung beziehe sich auf Bewußtseinsinhalte, welche die Tendenz haben, „im Bewußtsein weiter zu verharren“. Man muß also ein Minimum an Gedächtnisleistung voraussetzen, und weiter annehmen, das Erinnernte sei mit dem Erlebten identisch oder in bekannter Weise verschieden, wie Ach (S. 15) formuliert, man müsse „die Identität des perseverierenden Erlebnisses mit dem wirklich vorhandenen“ voraussetzen.

Schließlich ist als notwendige Voraussetzung zu erwähnen, daß die psychischen Gegebenheiten überhaupt bewußt und „bewußtseinsfähig“ sind. Die frühen Anwender der Introspektion gehen keinesfalls davon aus, alles Psychische sei bewußt oder könne bewußt gemacht werden. Dieser Ausgangspunkt hat jedoch nicht zu einer Theorie oder Kontroverse geführt, was unter welchen Umständen prinzipiell erlebt werden kann; die Grenzziehung wurde der Vp, dem VI und ihrer gemeinsamen alltäglichen und wissenschaftlichen Sprache überlassen.

2.3.4 Anmerkungen zu typischen Ergebnissen

Ergebnisse vorbehavioristischer Studien mit einer Methode der Selbstbeobachtung liegen oft auf drei verschiedenen Abstraktionsstufen vor: 1) als umfangreiche wörtliche Mitschrift, in der Regel wegen des Umfangs nicht oder nicht vollständig veröffentlicht, auch wegen ihrer Unvollständigkeit wohl nicht gerne aus der Hand gegeben, wobei sich Ach einiges von der Einführung des Phonographen verspricht 2) als kondensiertes Protokoll, deren Wiedergabe oft die Hälfte einer Publikation ausmacht und den Leser im interpretativen Nachvollzug zu 3) den theoretischen Schlußfolgerungen führen soll. Der Übergang von der einen zur anderen Ebene ist noch nicht systematisch kontrolliert; Prinzipien und Methoden etwa der Inhaltsanalyse waren damals noch nicht bekannt.

„Was freilich in der Frühzeit des experimentellen Arbeitens mit Selbstbeobachtungen ebenso wie bei manchen philosophischen Autoren vergessen wurde, ist der mittelbare Aussagewert von sprachlichen Äußerungen über innerseelische Vorgänge. Jede Aussage über ein Phänomen stellt ja nicht ohne weiteres die Abbildung einer Tatsache dar. Sie ist zunächst einmal eine verbale Reaktion auf eine Situation, die interpretiert werden muß wie irgendeine andere Reaktion.“ (Thomae 1960, S.32).

Thomae hat konsequent für Selbstbeobachtungen bei intraindividuellen Konflikten Verfahren der systematischen Inhaltsanalyse auf verbale Schilderungen von Erleben und Verhalten angewendet; Feger und Feger (1969 a, b) haben die Anwendung der Inhaltsanalyse auf erlebnisdeskriptives Material der Entscheidungsforschung weiterentwickelt. Um die nachfolgende Kritik verständlich zu machen, sei hier ein Teil eines typischen kondensierten Protokolls (Ach, S. 38 f.) wiedergegeben:

„Bei Jetzt wurde der Finger auf den Taster niedergedrückt mit dem Wissen, daß er niedergedrückt werden soll. Dann wurde die Blechplatte (Verschlußplatte des Kartenwechslers) fixiert und innerlich gesprochen „wird gleich kommen“ oder „jetzt kommts“, „jetzt kommts“ mit der Bedeutung, daß dort, wo fixiert wird, etwas (i.e. weiße Karte) eintreten wird. Dabei bestanden Spannungsempfindungen als sinnliche Begleiterscheinungen der Aufmerksamkeitskonzentration in den Augen, Stirn- und Schläfen, zuweilen auch in den Gesichtsmuskeln und in den Schultern, sowie ein Anhalten des Atems. Spannungsempfindungen in der Hand oder im Finger waren nur ausnahmsweise vorhanden. Trotzdem war in dem gesamten Spannungszustand das Wissen enthalten, daß sofort reagiert werden soll, ohne daß dies innerlich gesprochen wurde, oder sonst phänomenologisch repräsentiert war. Außerdem bestand die Bewußtheit, daß in kurzer Zeit das Erwartete eintreten, d.h. die Karte kommen wird, also neben der sonstigen Bestimmtheit des Erwartens auch eine zeitliche Komponente.

Die Erwartung selbst konzentrierte sich auf die kommende Karte, so daß diese im Mittelpunkt des gesamten Erlebnisses stand. Aber nur ausnahmsweise war dieselbe visuell gegeben (1. Versuch des 5. Tages), und auch hier war es nur „wie die Andeutung eines visuellen Streifens an der oberen Kante der Verschlußplatte, dessen Helligkeits-

qualität nicht als weiß zu bezeichnen war.“ Sonst war die weiße Karte nur als Bewußtheit im Erwartungsinhalt gegenwärtig d.h. Versuchsperson wußte, daß dort, wo sie fixierte, die weiße Karte erscheinen wird, ohne daß dieser auf die weiße Karte sich beziehende Vorstellungsinhalt anschaulich repräsentiert war.“

In diesem Protokoll erscheinen schon wesentliche theoretische Termini wie „Spannungsempfindung“, „Aufmerksamkeitskonzentration“, „Bewußtheit“, „Erwartungsinhalt“ und „Vorstellungsinhalt“. Es ist nicht durchgängig klar, für welche verbale Äußerungen der Vp sie jeweils stehen, und es ist nicht nachprüfbar, wie stark in der sich bisweilen über Wochen erstreckenden Interaktion zwischen VI und Vp die Anregungen des VI waren, entsprechende sprachliche Beschreibungen zu erzeugen. Pointiert formuliert wird die Vp zur Projektionsfläche für die theoretischen Vorstellungen des VI, und zwar so wenig kontrolliert, wie wir es uns heute kaum noch vorstellen können und gestatten würden.

2.4 Die behavioristische Kritik der „Introspektion“

Der Behaviorismus, den Watson (1913; 1968) von Anfang an selbst so taufte, verstand sich, ebenso von Anfang an, als Gegenwendung gegen eine mit Introspektion arbeitende Bewußtseinspsychologie. Das bekannte behavioristische Manifest von 1913 beginnt mit den Sätzen:

„Psychologie, wie sie der Behaviorist sieht, ist ein vollkommen objektiver, experimenteller Zweig der Naturwissenschaft. Ihr theoretisches Ziel ist die Vorhersage und Kontrolle von Verhalten. Introspektion spielt keine wesentliche Rolle in ihren Methoden . . .“ (1968, 13).

Watsons Vorbehalte waren insofern doppelter Art, als sie sich ineins gegen das Bewußtsein als Forschungsgegenstand der Psychologie und gegen die Introspektion als Forschungsmethode wandten, wobei die beiden Seiten dieser Kritik wie - in Watsons Sicht - auch ihre beiden Objekte einander bedingten. Es wäre heute nur mehr von historischem Interesse, Watsons oft durch Polemik vergrößerte Kritik der Introspektion in einen methodologischen Beitrag aufzunehmen. Die behavioristische Kritik als solche findet sich ohnehin differenzierter bei dem Protagonisten eines Radikalen Behaviorismus, B. F. Skinner (1953; 1963; 1974). Seine Kritik richtet sich nur gegen die Methode der Introspektion, nicht gegen das Bewußtsein. Als „private world within the skin“ gilt das Innere als ein kleiner Teil des gleichen Universums, das wir außen besser beobachten können. Aber für den inneren Teil gilt:

„Wir fühlen ihn und beobachten ihn auch in einem gewissen Sinn, und es wäre unsinnig, diese Informationsquelle bloß deswegen zu vernachlässigen, weil nie mehr als eine Person Kontakt mit einer inneren Welt aufnehmen kann. Nur

bedarf unser Verhalten bei dieser Kontaktnahme der Überprüfung“ (Skinner 1974, 24).

Berücksichtigt man den sozialen Ursprung unserer Selbstkenntnis und die Tatsache, daß etwas Fühlen und darüber Berichten zweierlei ist, dann ist auch die behutsame Verwendung von Berichten über „private Ereignisse innerhalb der Haut“ zu vertreten. über den „praktischen Nutzen von Berichten über die innere Welt, die gefühlt und introspektiv beobachtet wird“, heißt es, daß sie Anhaltspunkte ergeben (1) für vergangenes Verhalten und die Bedingungen, unter denen es stand, (2) für laufendes Verhalten und dessen Bedingungen und (3) für Bedingungen, die sich auf künftiges Verhalten beziehen“ (Skinner, a.a.O., 35).

Die Nutzenanwendung dieser Empfehlung findet sich in den entsprechenden mit Selbst-Beobachtung und Selbstkontrolle operierenden Techniken der (klinischen) Verhaltensmodifikation (vgl. hierzu Kanfer 1975; Braun 1978).

Für die Grundlagenforschung hingegen dominieren in Skinners eigenen Arbeiten die Bedenken gegenüber dem Einsatz introspektiver Methodik. Sie lassen sich dahingehend zusammenfassen, daß (1) introspektive Berichte nie genau genug sein können, weil die Entsprechung zwischen den privat bleibenden inneren Ereignissen und den sie (wenn überhaupt) begleitenden Umweltkontingenzen und dem sie kommentierenden (erklärenden oder beschreibenden?) verbalen Verhalten nie ganz verlässlich überprüft werden kann und zwar prinzipiell nicht; (2) die Berücksichtigung mentaler Ereignisse für eine *funktionale* Analyse des Verhaltens entbehrlich sei. In der funktionalistisch verstandenen Kausalkette ist es wichtig, das erste Glied (genetische und/oder Umweltbedingungen) zu kennen bzw. zu beherrschen, um - unter Außerachtlassen des mittleren mentalen Gliedes - das dritte Glied des manifesten Verhaltens vorhersagen bzw. modifizieren zu können. Die Kenntnis des zweiten Gliedes vermag ein gewisses Licht auf die Gesamtbeziehung zu werfen, sie aber nicht zu ändern (Skinner 1953, 35). Außerdem - und man mag dies als einen eigenen Einwand betrachten - fördert (3) die Berücksichtigung mentaler Ereignisse die Neigung und die Gefahr, allzusehnell Verhaltensweisen und -änderungen auf hypothetische innere Variablen zurückzuführen, statt weiter nach äußeren Ursachen zu forschen.

Wenn man ebenso abwägend wie Skinner die Vor- und Nachteile der Selbstbeobachtung darstellt, die Skinnersche und mit ihr überhaupt die behavioristische Kritik an introspektiven Verfahren beurteilt, kann man, wie es D.A. Lieberman (1979) getan hat, durchaus für eine (limitierte) Wiederzulassung der Selbstbeobachtung plädieren. Zur spät- bzw. nachbehavioristischen Rehabilitation der introspektiven Methodik vgl. auch Bakan (1959); Natsoulas (1970; 1978); Dikington & Glasgow (1967); Radford (1974) und White (1980); spe-

ziell zur historischen Rehabilitation ihrer Darstellung die - vor allem Boring (1953) korrigierende - Arbeit von Kurt Danziger (1980).

2.5 Die Technik des lauten Denkens

Was wir bei Kindern ohne Schwierigkeit beobachten und wobei wir auch uns selbst, wenn wir ganz allein sind, noch gelegentlich ertappen, ist ein mehr oder weniger fragmentarisches lautes Denken. Beim Lösen eines Problems hören wir den Betreffenden einzelne Lösungsschritte ankündigen, andere bereits getane rekapitulieren, gegenwärtige kommentieren usw.

Es war der Schweizer Denkpsychologe Edouard Claparède, der, unzufrieden mit der systematischen experimentellen Selbstbeobachtung der bisherigen Denkpsychologie, die von ihm „gesprochenes Denken“ (*reflexion parlée*) genannte Methode 1917 einführte (Claparède 1965). Karl Duncker übernahm sie dann 1926 (Duncker 1926; 1966).

„(Die Methode) besteht darin, jemanden die Lösung eines mehr oder minder schwierigen Problems aufzutragen... und diese Versuchsperson zu bitten, laut zu denken. Es handelt sich hier nicht um Introspektion, denn diese besteht aus der Analyse von Bewußtseinsprozessen, aus der Beschreibung ihrer Eigenart und ihrer Struktur. Hier handelt es sich einfach darum, die Denkschritte zu erzählen; es gilt zu beobachten, was das Denken tut, nicht, was es *ist*. Wenn Sie wollen, ist es eine behavioristische Methode, die den Ablauf des inneren Verhaltens zu bestimmen versucht. Vor der Introspektion hat sie den Vorteil, daß sie keine Spaltung der Versuchsperson verlangt: diese muß nicht zugleich denken und sich denken sehen“ (Claparède, a.a.O., 110).

Ähnlich argumentiert und operiert Duncker (1926, 664; 1966, 2), der die Methode unter Bezeichnungen des „thinking aloud“ und des „lauten Denkens“ bekannt gemacht hat. Sicher gibt es, wie eigene Empirie belegt, immer wieder mal Versuchspersonen, denen das Verbalisieren schwerfällt, die der Zwang zum Sprechen beim Denken irritiert. Und ebenso sicher gibt es keine Möglichkeit zu überprüfen, wieviele Gedanken ungeäußert bleiben. Trotzdem hat diese Methode, vor allem, wenn sie beim Problemlösen mit Verhaltensbeobachtung gepaart werden kann, in der modernen Denkforschung ihren festen Platz gefunden; man vergleiche etwa die Untersuchungen von Lüer (1973) und Dörner (1974) und die Behandlung der Methode als „weitverbreitete“ Datenquelle bei Ericsson & Simon (1980).

2.6 Phänomendeskription

Schon in den späteren Auflagen der Arbeiten von Wilhelm Wundt setzte sich allmählich der Begriff des Erlebens durch, zuerst noch gebunden als „Bewußt-

seinerlebnis“. Dieser neue bald und bis heute eingebürgerte Themenbegriff der Psychologie stand schließlich für alle Modalitäten unmittelbarer Erfahrung, ob Wahrnehmen, Vorstellen, Denken, Urteilen, Fühlen oder Wollen. Parallel zu dieser Entwicklung, sie wohl mitbedingend, hatte sich in der Phänomenologie ab der Jahrhundertwende eine neue Bewußtseinskonzeption durchgesetzt, deren wesentliches Bestimmungsstück Intentionalität war (s.O.; vgl. hierzu auch Gurwitsch 1966). An die Stelle eines „Behältnismodells“ des Bewußtseins, in dessen „Innerem“ man Inhalte und Akte (Prozesse) ansetzte, die man durch „Introspektion“ im engeren Sinne zu beobachten trachtete, trat das Modell eines „intentional“ auf die „Dinge selbst“ gerichteten Bewußtseins (bzw. Erlebens) (vgl. hierzu Graumann 1966; Nuttin 1955).

Das, was sich diesem Bewußtsein zeigte („Phänomen“), konnte der Deskription und weiteren Analysen unterzogen werden, so *wie es sich zeigte und rein in den Grenzen, in denen es sich zeigte*. Das methodische Postulat, das die Phänomendeskription betreibende Psychologie sich - unter dem Eindruck der Husserlschen Phänomenologie - verordnete, lautet in der erkenntniskritisch vielleicht anspruchsvollsten Fassung bzw. Forderung:

„Das Vorgefundene zunächst einfach hinzunehmen, wie es ist; auch wenn es ungewohnt, unerwartet, unlogisch, widersinnig erscheint und unbezweifelten Annahmen oder vertrauten Gedankengängen widerspricht. Die Dinge selbst sprechen zu lassen, ohne Seitenblicke auf Bekanntes, früher Gelerntes, ‚Selbstverständliches‘, auf inhaltliches Wissen, Forderungen der Logik, Voreingenommenheiten des Sprachgebrauchs und Lücken des Wortschatzes. Der Sache mit Ehrfurcht und Liebe gegenüberzutreten, Zweifel und Mißtrauen aber gegebenenfalls zunächst vor allem gegen die Voraussetzungen und Begriffe zu richten, mit denen man das Gegebene bis dahin zu fassen suchte“ (Metzger 1954, 12).

Nur um einem gegenüber einer deskriptiv vorgehenden oder gar phänomenologisch orientierten Psychologie häufig aktualisierten Vorurteil entgegenzuwirken, sei der obigen Forderung noch mit Wolfgang Metzger die Anmerkung beigelegt, daß dieses Deskriptionsgebot keinen Verzicht auf weiterführende Hypothesenbildung und deren experimentelle Prüfung bedeutet, wohl aber, „daß es sinnlos ist und zu Fehlansätzen führen muß, wenn man zu Annahmen und Untersuchungen über Ursachen und über Wirkungen des zunächst Gegebenen übergeht, ohne dieses überhaupt recht zu kennen“ (Metzger a.a.O., 13).

Modellbeispiele für eine derartige Phänomendeskription, die hier aus Raumgründen schlecht wiedergegeben werden können (sie sind, weil differenziert, entsprechend umfangreich), finden sich in der phänomenologischen Literatur etwa bei W. Schapp (1976), in der psychologischen Literatur etwa bei David Katz (1911; 1929); im übrigen sind viele der heute klassisch genannten Arbei-

ten der Gestaltpsychologen seit Wertheimer (1912) und Köhler (1921) mit Hilfe solcher Phänomendeskription zu ihren Ergebnissen gekommen. Demgegenüber ist die neuere Psychologie arm an derartigen Dokumenten reiner Erlebnisdeskription (vgl. aber Gibson 1950; deutsch 1973).

2.7 Behavioristische Selbstwahrnehmung

Eine der neueren Darstellungen der behavioristischen Auffassungen gibt Bern (1972, S. 2), der die folgende Zusammenfassung als die zentralen Annahmen seiner Theorie der Selbstwahrnehmung bezeichnet:

„Individuals come to „know“ their own attitudes, emotions, and other internal states partially by inferring them from observations of their own overt behavior and/or the circumstances in which this behavior occurs. Thus, to the extent that internal cues are weak, ambiguous, or uninterpretable, the individual is functionally in the same position as an outside observer, an observer who must necessarily rely upon those same external cues to infer the individual's inner states.“

Diese Position geht im wesentlichen zurück auf die radikal-behavioristische Analyse sog. „private events“ durch Skinner (1945, 1953, 1957). Bern behandelt das Thema weiter unter der aufschlußreichen Kapitelüberschrift „The ontogeny of self-attributions“. Um innere Zustände richtig beschreiben zu können, muß ein Kind dies lernen, und zwar von jemandem, der es Namen für diese Zustände lehrt und ihm beim Unterscheiden ähnlicher Zustände hilft. Dabei ergibt sich das Problem, daß dieser Lehrer, der „außenstehende“ Beobachter, den Zeitpunkt und die Gelegenheit feststellen muß, wo beim Kind der „kritische innere Reiz“ auftritt, z.B. Schmerz, wenn es sich den Kopf stößt. Die Beschreibung „das tut weh“ mag dann in das Reaktionsrepertoire des Kindes aufgenommen werden, wenn beispielsweise die Mutter sagt: „Weine nicht, ich weiß ja, daß es dir weh tut.“ Das Kind kann diese Beschreibungs-Reaktion generalisieren auf die Wirkung einer größeren Zahl schmerzzeugender Reize. Am Anfang stand - nach Skinner wie Bern - ein beobachtbarer äußerer Reiz (Kopf stoßen) und vielleicht eine beobachtbare Reaktion (Weinen). Es ist dann rigoros und konsequent anzunehmen, „. . . that we have virtually no knowledge at all until we have been explicitly trained. Internal identifications that we have not been taught remain internal identifications that we cannot make.“

Die Überprüfung der Theorie, insbesondere die Ableitung spezifischer, empirisch prüfbarer Hypothesen, ist noch nicht weit vorangekommen. Berns (1965; 1966) hierzu angestellte Experimente sind schwierig zu bewerten und setzen sich im wesentlichen mit der Dissonanztheorie auseinander. Wir wenden uns daher einigen Unterschieden zu, die Bern (1972) zwischen Selbstwahrnehmung und interpersonalen Wahrnehmung herausstellt. Die erste Unterscheidung

wird als die zwischen *insider versus Outsider* getroffen. Dem Selbstbeobachter des eigenen Erlebens stünden Reize zur Verfügung, die dem Fremdbeobachter nicht verfügbar seien, und auch zwischen diesen Reizen könne diskriminiert werden, wenn auch deshalb relativ schlecht, weil die Gemeinschaft derer, die die gleiche Sprache sprechen, nur begrenzt imstande ist, solche Differenzierungen einzuüben. Die zweite Unterscheidung ist die zwischen *intimate versus stranger*. Für den Selbstbeobachter steht die Kenntnis der eigenen Vergangenheit zur Verfügung, die für den Außenstehenden nicht als Bezugssystem dienen kann. Ihm fehlen damit Anhaltspunkte, um die inneren und äußeren Reize und Reaktionen zu bewerten. Weiter wird abgehoben auf die besonderen und möglicherweise verschiedenen Interessenlagen von *self versus other*, die zu Verzerrungstendenzen führen könnten. Schließlich wird die Perspektive des Selbstbeobachtenden als Handelnden gegen die nur observierende des Fremdbeobachters gestellt - actor versus *observer*, und aus den unterschiedlichen Perspektiven können unterschiedliche Aspekte der Situation relevant werden.

2.8 Neuere Untersuchungen über bildhafte Vorstellungen

Das in den letzten Jahrzehnten verstärkte Interesse an kognitiver Psychologie und an Phänomenen wie den bildhaften Vorstellungen (mental images oder imagery), die in der Tradition der älteren Psychologie als Bewußtseinsstatsachen zu bezeichnen wären, könnte vermuten lassen, man hätte Verfahren zur Erfassung von Erleben in diesem Zusammenhang diskutiert, oder gar gezielt untersucht und verbessert. Uns interessiert hier nicht die Rolle, die bildhafte Vorstellungen in verschiedenen kognitiven Theorien spielen, sondern die Methoden zu ihrer Analyse. Richardson (1980) berichtet in seiner Monographie drei Vorgehensweisen, die fast die gesamte Forschung beschreiben:

- (1) Vpn werden instruiert, sich etwas bildhaft vorzustellen, wenn sie bestimmte Aufgaben - meist des Lernens und Erinnerns - im Experiment ausführen.
- (2) Den Vpn wird unterschiedliches Reizmaterial vorgelegt und man erhebt, in welchem Ausmaß das Material bildhafte Vorstellungen erzeugt, und ob dieses berichtete Ausmaß mit Leistungsvariablen zusammenhängt. Meist wird geprüft, ob Material, das mehr bildhafte Vorstellungen hervorruft, besser behalten wird.
- (3) Man vergleicht Vpn, die sich in der Lebhaftigkeit der von ihnen als erlebt berichteten bildhaften Vorstellungen unterscheiden, nach ihren Leistungen in Tests, welche die Fähigkeit bei räumlichen Manipulationen erfassen, z.B. bei der vorgestellten Rotation vorgegebener oder vorgestellter Figuren. Reaktionszeit ist eine hier typische abhängige Variable.

Um die in diesem Forschungsbereich verwendeten Methoden zu diskutieren, schildern wir das Vorgehen anhand der Arbeiten von Shepard (Übersichtsarti-

kel 1978). Shepard untersucht bildhafte Vorstellungen von realen Objekten, und zwar vorgestellten gegenüber als Vorlage im Versuch vorgegebenen, die nicht ambivalent sind, also z.B. nicht mehrdeutige optische Figuren oder uneindeutige, TAT-ähnliche Reizvorlagen. Auf diese Weise versucht er, die Probleme zu umgehen, die sich daraus ergeben, daß subjektive Deutungen - man sieht etwas als etwas - bei einigen Reizvorlagen die entscheidende Rolle für interindividuelle Varianz spielen. Die Grundannahme dieses Versuches (Shepard & Chipman, 1970) wie zahlreicher anderer besteht darin, eine Äquivalenz zwischen Wahrnehmung und Vorstellen zu vermuten, und zwar „... a more abstract or ‚second order‘ isomorphism in which the functional relations among objects as imagined must to some degree mirror the functional relations among those same objects as actually perceived.“ (Shepard 1978, S. 131). Dann müßten - eine weitere Annahme - Vpn Fragen nach der Ähnlichkeit von Komponenten des vorgestellten Bildes in etwa so beantworten wie Fragen nach der Ähnlichkeit von Komponenten des visuell vorhandenen Bildes. Als typisches Ergebnis werden statistisch nicht unterscheidbare Matrizen von Ähnlichkeitskoeffizienten berichtet, und man schließt, „... the subjects were performing very similar mental processes in the perceptual and imaginal conditions“ (S. 132). Da bei den Vpn ohne visuelle Vorgabe eine bildhafte Vorstellung nicht experimentell hergestellt wird, ist es denkbar, daß sie ohne solche Vorstellungen zu ihrem Verhalten kamen, daß sie etwa den demand characteristics des Experimentes entnehmen, wie sie reagieren sollten (vgl. Mitchell & Richman 1980). Forscher wie Finke & Kosslyn (1980) reagieren auf diesen Einwand mit dem Bemühen, experimentelle Situationen zu erfinden, in denen Vpn nicht wissen können, was die „richtigen“ Reaktionen wären, weil diese zu kompliziert sind.

Andere Versuchspläne sind gegenüber dieser Kritik noch anfälliger, wenn beispielsweise eine Gruppe von Vpn aufgefordert wird, sich vor der Reizdarbietung eine Vorstellung dieses Reizes zu machen, und diese Gruppe mit einer anderen ohne diese Instruktion verglichen wird (z.B. Shepard & Metzler 1971). Während in der zuvor erwähnten Untersuchung Ähnlichkeitseinstufungen als abhängige Variable verwendet wurde, also verbale Berichte über Ergebnisse kognitiver Prozesse, werden im zweiten Experiment, wie auch sonst häufig, Reaktionszeiten verwendet, also Verhaltensbeobachtungen als abhängige Variablen. Es darf als charakteristisch für die heutige kognitive Psychologie, und als Folge der behavioristischen Kritik am Introspektionismus gelten, daß Annahmen über Erleben, über bewußte Prozesse, an ihren vermuteten Auswirkungen auf Verhalten überprüft werden; nur selten greift man auf Selbstbeschreibungen zurück. Kognitive Vorgänge werden eher als intervenierende Variablen denn als hypothetische Konstrukte aufgefaßt.

In einer dritten Art von Versuchsplänen bemüht man sich, die Vp zu direkten Operationen an ihren bildlichen Vorstellungen zu veranlassen, und diese Ope-

rationen sollen die gleichen sein wie an vorgegebenem Material. Podgorny & Shepard (1978) beispielsweise legten ihren Vpn in der visuellen Versuchsbedingung den in Abb. 1 A gezeigten Blockbuchstaben F vor, in der Vorstellungsbedingung das leere Raster B in Abb. 1, zugleich mit der Instruktion, sich den mit Hilfe eines zuvor gelernten Codes beschriebenen Buchstaben F in diesem Raster vorzustellen. Wenn die Vp ihre volle Bereitschaft signalisierte, wurde ihr als Probereiz ein Punkt im Raster, Teil C in Abb. 1, z.B. mit der Frage dargeboten, ob der Probereiz auf der Figur läge oder nicht. Außer der Antwort wurde auch die Reaktionszeit erfaßt. Auch bei komplizierten Vorgaben stimmten die Ergebnisse der visuellen und der Vorstellungs-Bedingung überein, so waren beispielsweise die Reaktionen kürzer, wenn mehrere Punkte eines zusammengesetzten Probereizes auf die Figur fielen als wenn sie auf unterschiedliche Teile des Rasters fielen. Anscheinend wurden die Vorstellungen so erzeugt, daß gleichartige Operationen, jedenfalls solche mit vergleichbarem Zeitbedarf, bei den gleichen Aufgaben durchgeführt wurden.

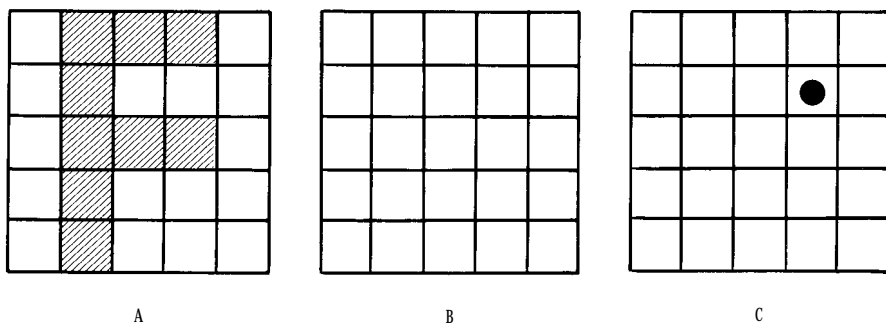


Abb. 1: Vorlagen im Versuch von Podgorny & Shepard, 1978

Beim nächsten Typus von Versuchsplan sollten laut Instruktion Operationen mit den Vorstellungen selbst vorgenommen werden, beispielsweise bei Shepard & Metzler (1971) eine Rotation dreidimensionaler Objekte im „Vorstellungsraum“. Gestützt auf mehrere solcher Experimente kommt Shepard (1978, S. 134) zu dem Schluß: „... we have established, more directly, that the intermediate states of the internal process do indeed have a one-to-one correspondence to intermediate orientations of the external Object. Our results, in fact, show that there is actually something rotating during the course of a mental rotation - namely, the orientation in which the corresponding external stimulus, if it were to be presented, would be most rapidly discriminated from other possible stimuli.“

Zweifellos stellen Ergebnisse wie diese und ähnliche jemanden, der sie ohne das Auftreten *bildhafter* Vorstellungen anzunehmen erklären will, vor eine

schwierige Aufgabe. Dennoch ist in allen diesen Untersuchungen der theoretisch interessierende Zustand, methodisch gesprochen: die unabhängige Variable, nicht experimentell hergestellt worden; der Erfolg der experimentellen Manipulation, die oft lediglich aus einer Instruktion an die Vpn besteht, ist kaum oder meistens gar nicht kontrolliert worden. Wenn auch im letzten Jahrzehnt die Annahmen über die Natur der untersuchten kognitiven Prozesse und Strukturen deutlicher geworden sind, so gilt dies nicht für die den Methoden impliziten Annahmen, die eine Methode bei ihrer Anwendung zu einer geheimen Theorie des Gegenstandsbereiches machen, zu dessen Erforschung sie gerade eingesetzt werden.

2.9 Methoden der Metakognitionsforschung

Die „kognitive Wende“ oder gar „kognitive Revolution“, die seit einigen Jahren selbst aus den Texten der ernsthaftesten psychologischen Grundlagenforscher und Theoretiker in die Augen springt (und an die „behavioristische Wende“ nach 1913 erinnert), besagt vorerst nur, daß tatsächlich der kognitive Wortschatz eine unwahrscheinliche Blüte erlebt. Weniger klar ist, inwieweit der neue Diskurs eine „Wende“, d.h. eine Abkehr vom „Behaviorismus“, symbolisiert. Zumindest steht das, was unter dem Zeichen eines Kognitivismus, zumeist auf dem Gebiet der „Artificial Intelligence“, allgemein der Informationsverarbeitungsmodelle, geleistet wird, nicht unbedingt in Widerspruch zu elaborierten Mediationsmodellen des Neobehaviorismus. Auch von einer nennenswerten Wiederbelebung der bewußtseinstheoretischen Diskussion kann keine Rede sein. Insofern steht auch die Weiterentwicklung der erlebnisdeskriptiven Methodik nicht unbedingt auf dem Programm der „Kognitiven Psychologie“.

Eine Ausnahme läßt sich jedoch erkennen: War die (iterative) Reflexivität des *cogito me cogitare* . . ., des ‚Ich weiß, daß ich weiß‘, auch ein jahrhundertealtes Thema philosophischer Reflexion, die Psychologie erreichte das Thema des Wissens über Wissen nur da, wo sich eine phänomenologische Einstellung, in der Bewußtsein thematisiert wird, realisierte. Das hat sich nun insofern geändert, als jetzt mehr und mehr Psychologen die alte Erkenntnis, daß das Tun eines Subjekts vom Bewußtsein dieses Tuns begleitet sein kann, sozusagen wiederentdecken und vor allem für Entwicklungs- und Pädagogische Psychologie fruchtbar zu machen suchen.

Um nicht in die nach wie vor noch tabuisierte Bewußtseinsterminologie der älteren Psychologie verstrickt zu werden, ist schlicht die Rede von *Metakognitionen* (metamemory, metalearning, metaattention), auch wenn mit (schlecht definierbaren) „Kognitionen über Kognitionen“ (Meichenbaum et al. 1979) doch nur „knowing about knowing“ (Brown 1978) gemeint ist. Immerhin

dürfte dieser Bereich der menschlichen Reflexivität, des Wissens, daß wir wissen und, günstigenfalls, wie wir lernen, denken, uns erinnern, am ehesten von allen kognitiven Forschungsgebieten auf den Einsatz erlebnisdeskriptiver Verfahren angewiesen sein.

Wiederum hängen aber der Einsatz und die Verwendung der Methoden ab von der Verwendung (Bedeutung) der das Feld bezeichnenden Konzepte.

Wenngleich der Begriff bzw. die Vokabel *Metagedächtnis* am Anfang dieser neueren Entwicklung stand, muß man ihn wohl heute dem Konzept der *Metakognition* zuordnen:

„Metakognition bezieht sich auf das Wissen eines Menschen über seine eigenen kognitiven Prozesse und Produkte oder was damit in Beziehung steht, wie lernrelevante Eigenschaften von Informationen oder Daten. So betreibe ich z.B. Metakognition . . ., wenn ich bemerke, daß ich mehr Schwierigkeiten habe, A als B zu lernen; wenn mir aufgeht, daß ich C nochmal prüfen sollte, ehe ich es als Tatsache akzeptiere; wenn mir einfallt, daß ich besser jede einzelne Alternative in den Multiple-choice-Aufgaben durchprüfe, ehe ich entscheide, welche die richtige ist; wenn ich das Gefühl habe, ich sollte mir D lieber notieren, damit ich es nicht vergesse . . .“ (Flavell 1976, 232).

Rein von der Funktion her wird Metakognition primär als Selbstkontrolle und -Steuerung verstanden; als „*monitoring*“, was nach Flavell (1979, 2) soviel heißt wie „keeping track of how it is going and taking appropriate measures whenever it needs to go differently“. Unabhängig davon kann man das „metakognitive“ Wissen primär als *Inhalt* (stored contents) oder primär als Prozeß bzw. *Aktivität* auffassen (Cavanaugh & Perlmutter 1980).

Je nach Akzentsetzung fallen denn auch die methodischen Präferenzen aus: beim inhaltsorientierten Vorgehen Interviews und Fragebogen, bei Aktivitätsorientierung verbale Protokolle, Verhaltensbeobachtung, „feeling-of-knowing“-Technik.

Als jüngster Versuch in der Geschichte der Psychologie, wenigstens des (reflexiven) Wissens unseres Bewußtseins habhaft zu werden, verdient seine Methodik besondere Aufmerksamkeit.

Doch wird derjenige, der die Entwicklung oder den Einsatz einer neuartigen Methode erwartet, vorerst enttäuscht. Wie schon gelegentlich in der Diagnostik praktiziert, wird dieser Mangel durch eine „Batterie“ von Verfahren kompensiert. Dabei stellen Interview und Fragebogen die bisher verbreitetsten Verfahren dar, deren Schwächen, im Prinzip bekannt, neu diskutiert werden (Adair & Spinner 1979; Meichenbaum & Butler 1980). Bei dem von Hart (1965; 1966; 1967) entwickelten ‚Feeling-of-knowing‘-Verfahren werden Vpn, die sicher sind, die Lösung eines Problems bzw. die Antwort auf eine Frage

schon „auf der Zunge“ zu haben, sie nur nicht aussprechen zu können („tip-of-the-tongue-Phänomen“), zu Aussagen über die entsprechenden Items veranlaßt, die dann mit späteren Aussagen darüber verglichen werden (kritisch hierzu: Cavanaugh & Perlmutter 1980). Problematisch scheint auch die Verwendung der Reaktionszeit für eine Frage als Indikator der Gewißheit der Vp bezüglich ihres Wissens zu sein (Lachman & Lachman 1980). Nahe liegt, daß zur Identifikation von Metakognitionen auch die Methode des lauten Denkens wieder aufgegriffen wird. Sie wird ergänzt durch eine Art Nachfaßtechnik (probe technique), bei der die Vpn nach jeder Aufgabe nach den Strategien und Hypothesen gefragt werden, die sie bei der Bearbeitung benutzt haben. Ähnlich hatte schon Giorgi (1967), um die Gleichgewichtigkeit von „experimental data“ und „experiential data“ zu demonstrieren, nach einem traditionellen Experiment zum seriellen Lernen seine Vpn in einer postexperimentellen Befragung zur subjektiven Schwierigkeit u.ä. vernommen und dadurch erst Zugang zu den Lösungsansätzen (und zur sinnvollen Interpretation seiner Daten) gefunden. Es besteht die Hoffnung, daß durch das jüngste Interesse an Metakognitionen die methodisch noch entwicklungsfähige Kombination von Verhaltens- bzw. Leistungs- und Erlebnisdaten neue Impulse erfährt.

3. Aktuelle Probleme der Verhaltensbeobachtung

Anders als im Artikel über die wissenschaftliche Beobachtung (Feger, in diesem Band), in dem allgemeine methodische Probleme der wissenschaftlichen Beobachtung besprochen wurden, wenden wir uns hier Ansätzen zu, die versuchen, Beobachtungsverhalten mit Hilfe psychologischer Begriffe und Theorien zu beschreiben und zu erklären. Schließlich ist Beobachten menschliches Verhalten, und somit Bestandteil des Gegenstandsbereiches unseres Faches. Wir befassen uns zunächst mit der Frage, wie der Gegenstand einer psychologischen Verhaltensbeobachtung zu bestimmen sei, dann mit der teilnehmenden Beobachtung, mit den Regeln, nach denen Beobachter ihre Beobachtungseinheiten festlegen, und schließlich mit den Zusammenhängen zwischen Beobachtung, Gedächtnis und verbalem Bericht über das Beobachtete.

3.1 Der Gegenstand psychologischer Verhaltensbeobachtung

Wir stellen zunächst die Frage, was gegenwärtig typischerweise *Untersuchungsgegenstand psychologischer Beobachtungsstudien* ist und wie er genau anzugeben sei, wobei wir der Übersicht von Longabaugh (1980) weitgehend folgen. Longabaugh unterscheidet zwei Brennpunkte der Analyse: Was zeigt sich als Verhalten, und *wie* zeigt es sich:

„The study of ‚what‘ is the study of what it is actors communicate to one another: both what is intended and encoded and what is received or decoded. The study of ‚how‘ is the study of how the content is expressed.“ (p. 64).

Die Untersuchung der Erscheinungsformen hat sich in den letzten Jahrzehnten stark ausgebreitet und spezialisiert, wobei sich zwei Schwerpunkte gebildet haben: Die Analyse des Sprechens und der nichtverbalen Kommunikation (z.B. Duncan 1969) und der Bewegungen des Körpers (z.B. Birdwhistell 1970, Schefflen 1975).

Neben Erscheinungsform und Inhalt werden, um den Untersuchungsgegenstand genau zu bestimmen, „actors, targets, and settings“ anzugeben sein. Man wird also diejenigen bestimmen müssen, von denen das Verhalten ausgeht, diejenigen, auf die es gerichtet ist, und die Umstände schildern müssen, unter denen das Verhalten verwirklicht wird. Longabaugh führt ein Klassifikationssystem von Lambert (1960) weiter, in dem für Verhaltensträger, Verhaltensziele und Verhaltensumwelt unterschieden wird, ob sie als einmalig und spezifisch festgelegt oder als Klasse umschrieben oder unspezifiziert gelassen werden. Für Verhaltensträger beispielsweise bedeutet die erste Kategorie, daß eine Einzelfallstudie durchgeführt wird, die zweite Kategorie, daß beispielsweise zwölfjährige Mädchen in Bayern untersucht werden, die dritte Kategorie, daß Ergebnisse ohne Angabe des Bereichs der Verallgemeinerbarkeit mitgeteilt werden. Als wichtig unterstreicht nun Longabaugh, daß auch Verhaltensziele und settings nach dem Grad der Spezifität charakterisiert werden können und müssen: „Because of the variable and sometimes overwhelming effects of targets and settings on behavior, researchers of behavior in naturalistic settings cannot let any of these elements go unidentified.“ Dabei sollte die Charakterisierung nach jenen Merkmalen erfolgen, die als für das beobachtete Verhalten als wesentliche Ursachen, Bedingungen, Voraussetzungen etc. angesehen, oder - besser noch - aus einer Theorie des Beobachteten *abgeleitet* werden. Dann ergibt sich als weiterer Vorteil neben der eindeutigen Abgrenzung des Aussagenbereichs die Möglichkeit, das Beobachtete umzuklassifizieren, insbesondere dann, wenn in einer Bedingungskombination mehr Heterogenität als erwartet beobachtet wurde.

Die Beschreibung der Umwelt kann nach Longabaugh aus drei Perspektiven erfolgen: erstens als die physikalische Beschreibung der Objekte und des Terrains (z.B. Barker 1968, Whiting & Whiting 1975), zweitens als Beziehung zwischen der beobachteten Person und ihrer Umwelt, wobei die *relevanten* Aspekte der Umwelt - zusammengefaßt als Situation des Beobachteten - aus der Kovariation von Verhalten und Umweltvariablen erschlossen wird (z.B. Skinner 1953), und schließlich drittens als die Bedeutung, die einer Gelegenheit, einem Ort, einem Anlaß von den typischen Benutzern, Anwesenden und Teilnehmern gemeinsam und übereinstimmend zugeschrieben wird (z.B. Barker & Schoggen 1973).

3.2 Analyse des Beobachters als Meßinstrument

3.2.1 Die Ermittlung von „Fehlern“

Beobachter werden häufig dann eingesetzt, wenn es keine Meßinstrumente gibt, um das zu erfassen, woran der Wissenschaftler interessiert ist. Der Beobachter übernimmt dann die Funktion eines Meßinstruments:

„When a person makes a perceptual judgment, he acts as a kind of measuring instrument, and the theory of perceptual judgments is an attempt to understand how these measurements are produced.“ (Greeno 1968, S. 1).

Der Gedanke liegt nahe, man müsse die Eigentümlichkeiten, Unvollkommenheiten, „Fehler“ dieses menschlichen Meßinstrumentes kennen, um sie bei Schlüssen aus den Beobachtungsergebnissen berücksichtigen zu können. Aus anderer Perspektive geben diese Fehler Aufschluß über den Menschen als kognitives System, das zu analysieren Aufgabe der Allgemeinen Psychologie ist. Im ersten Fall interessiert nach wie vor primär das Beobachtete, und die Fehleranalyse ist in Gefahr, ein Inventar von Kuriosa zu bleiben; im zweiten Fall interessiert der Beobachter als solcher. Zwischen diesen beiden Polen schwankten die Auffassungen in der Psychologiegeschichte, wobei der Fehlerauffassung historisch früher nachgegangen wurde, und zwar besonders dann, wenn in einem bestimmten, relativ umgrenzten Forschungsbereich - etwa bei der Personenbeschreibung in der Diagnostik und Sozialpsychologie, als Erwartungsfehler in der Psychophysik oder Kontexteffekt in der Wahrnehmungsforschung - sich solche Unregelmäßigkeiten zeigten, die mit dem umgrenzten Bereich anscheinend nichts zu tun hatten. Die Fehler scheinen relativ unabhängig von Inhalt des jeweils Wahrgenommenen und deshalb Merkmale von Beobachtern allgemein zu sein.

Guilford (1954, S. 278ff.) gibt eine Übersicht der bis dahin häufiger untersuchten Fehlerarten: der Fehler der zentralen Tendenz (error of central tendency), der Fehler der Milde (leniency) oder der Großzügigkeit (generosity error), logische Fehler, verzerrende Effekte der zeitlichen Reihenfolge: stärkerer Einfluß der zuerst (primacy-) oder zuletzt (recency effect) aufgenommenen Elemente einer Serie, und besonders der Halo-Effekt. Man könnte Eigentümlichkeiten im Umgang mit bestimmten Urteilsprachen hinzunehmen, etwa die bevorzugte Verwendung bestimmter Kategorien einer Rating-Skala, oder Einflüsse, die von anderen Zielen und Absichten des Beobachters ausgehen als gerade zu beobachten, beispielsweise die Wirkungen der sozialen Erwünschtheit von Reaktionen.

Von „Fehlern“ kann man nur sprechen, wenn es Diskrepanzen zwischen den Beobachtungen und Resultaten anderer, unabhängiger methodischer Zugänge zum gleichen Gegenstand gibt. Und die Gültigkeit und Verlässlichkeit dieser

Ergebnisse anderer Methoden muß hinreichend gesichert sein. Betrachtet man die erwähnten Fehlerarten in dieser Hinsicht näher, so erscheint die Problemlage nicht einheitlich und für die Annahme, es handele sich um Fehler, nicht günstig. Beispielsweise wurden Fehler deshalb angenommen, weil Urteile verschiedener Beobachter untereinander nicht übereinstimmten und man systematische Einflußquellen ermitteln konnte - z.B. beim Mildefehler um so größeres Wohlwollen, je bekannter der Beurteilte dem Beurteiler war. Oder Beobachtungsleistungen wurden als Fehler bezeichnet, weil sie mit Selbstaussagen der Beobachteten nicht übereinstimmten; Selbstaussagen sind sicher nicht garantiert valide. Beim Vergleich mit physikalischen Meßwerten, die an den beobachteten Phänomenen gewonnen wurden, ist das Fehlerkonzept in dem Augenblick überholt, wo man Wahrnehmung als aktive Herstellung einer psychischen Wirklichkeit durch den Organismus begreift. Wir wählen daher im folgenden exemplarisch einige Ansätze aus, die primär am Beobachter, und dann an seinen Urteilen interessiert sind.

3.2.2 Der Einfluß von semantischen Gedächtnisstrukturen auf Verhaltensbeschreibungen

D'Andrade (1974) untersucht die Auswirkungen von langfristigem (länger als 10 Min.) Gedächtnis auf Urteile über Verhalten. In einer früheren Arbeit (1965) zeigte er, daß Urteile über die semantische Ähnlichkeit von Eigenschaftsbezeichnungen einerseits, Verhaltenseinschätzungen andererseits nahezu die gleiche Struktur der Interkorrelationen aufwiesen. Die Isomorphismus-Hypothese deutet diesen Befund so, die Ähnlichkeit der Eigenschaftsnamen bestehe, weil sie ein Niederschlag der Erfahrungen über den tatsächlichen, beobachtbaren Zusammenhang von Verhaltenseigenschaften sei. Hingegen interpretiert die Verzerrungshypothese: Semantische Zusammenhänge verfälschen Urteile über Verhalten so, daß sie in der Weise interkorrelieren, wie es der semantischen Ähnlichkeit entspricht.

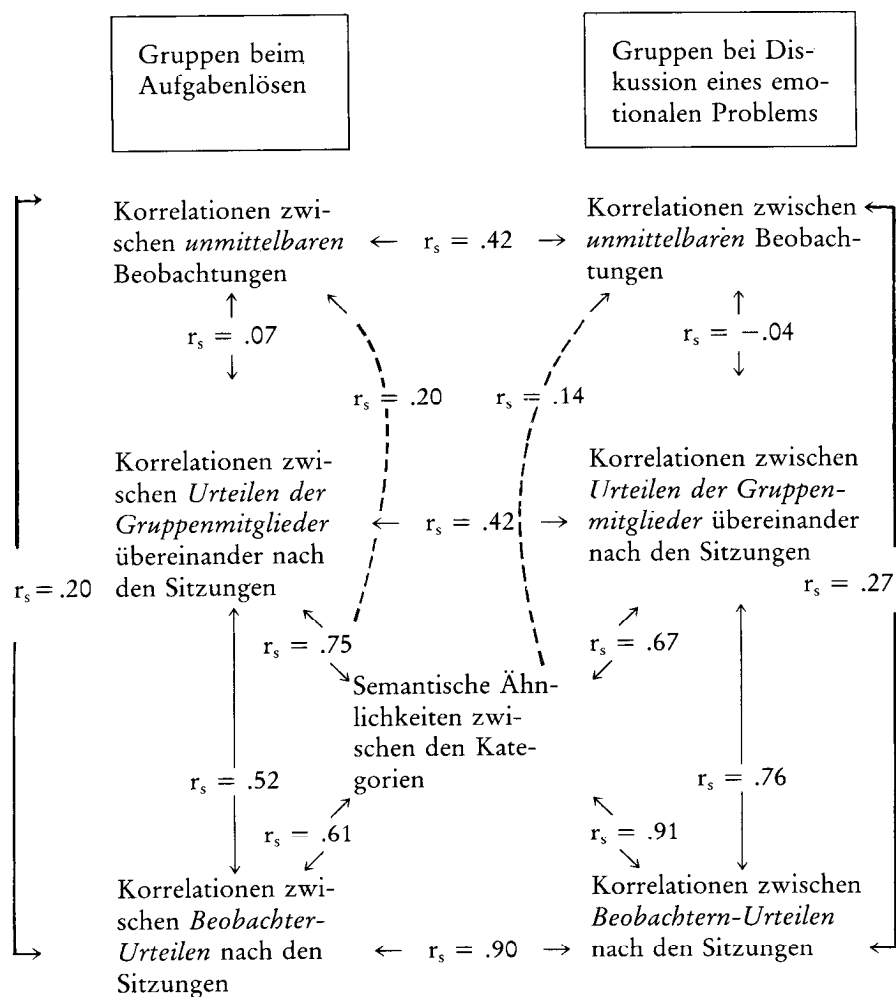
Zwischen beiden Hypothesen kann man nach D'Andrade entscheiden, wenn man über zweierlei Daten verfügt: unmittelbare, detaillierte Verhaltensbeobachtungen und Beschreibungen nach Ablauf eines längeren Zeitintervalls.

„If the observer's memory-based ratings showed a very different pattern of correlations from that found for the data based on the actual behavior of the subjects (but a pattern similar to judgments of semantic similarity), it would be reasonable to reject the isomorphism hypothesis and to consider the systematic-distortion hypothesis supported.“ (D'Andrade 1974, S. 162).

D'Andrade untersucht diese Frage anhand der Studien von Borgatta et al. (1958) und Mann (1959), deren Daten er Einschätzungen seiner eigenen Vpn hinzufügte, welche die semantische Ähnlichkeit zwischen den (in Anlehnung

an Bales formulierten) verhaltensdeskriptiven Kategorien eingestuft hatten. Bestimmt wurde dann die Ähnlichkeit zwischen den Korrelationsmatrizen, indem Spearmans rho zwischen den entsprechenden Korrelationen in je zwei Matrizen berechnet wurden. Die Ergebnisse, die sich auf die Mann (1959)-Studie beziehen, seien in Tab. 1 wiedergegeben. Sie stützen nach D'Andrade eindeutig die Verzerrungs-Hypothese. Im Gedächtnis wurden demnach Beobachtungen bereits bestehenden, semantischen Strukturen angeglichen; allen-

Tabelle 1: Vergleich von Korrelationsmatrizen mit Hilfe von Spearmans rho (= D'Andrade 1974, S. 176).



falls könnte man Beschreibungen trauen, die unmittelbar nach der Beobachtung entstehen. (Borg & Bergermaier, 1979, haben in anderem Zusammenhang die Problematik solcher Matrizenvergleiche herausgestellt.)

Mit der Schilderung der Arbeiten von D'Andrade sind wir in eine Kontroverse eingestiegen, die seit Beginn der wissenschaftlichen Psychologie zunächst aus der Perspektive der Beurteilungsfehler geführt wurde: Wenn Beobachter Verhalten beschreiben und (angeblich) verhaltensbezogene Schätzurteile abgeben, interessiert oft nicht nur die berichtete Ausprägung eines Merkmals, sondern der Zusammenhang zwischen mehreren Merkmalen. Und man vermutet, Schätzungen des Zusammenhanges von Merkmalen könnten verfälscht sein, weil die Beurteiler schon ihre Vorstellungen über den Zusammenhang der Merkmale hätten, und diese Vorstellungen gingen in noch näher zu bestimmender Weise in die Schätzungen ein.

Die Thematik wurde erstmals in den Forschungen zum *Halo-Effekt* angesprochen (Wells 1907, Thorndike 1920, Rugg 1922; s. auch Rudinger & Feger, 1970). Unter Halo-Effekt versteht man einen überhöhten Zusammenhang zwischen (beurteilten) Merkmalen, verursacht durch den Gesamteindruck über den Merkmalsträgern und beim Beurteiler bestehende Annahmen über die Beziehungsstruktur zwischen Merkmalen im allgemeinen. Hat ein Beurteiler einen insgesamt positiven Eindruck von einer Person gewonnen, so werde er dazu neigen, sie relativ günstig bei von ihm als positiv bewerteten, miteinander zusammenhängenden Eigenschaften einzuschätzen. Korreliert man nun die Urteile zu zwei Merkmalen (in der Literatur: der gleichen oder verschiedener Personen, von einem oder mehreren Beurteilern), so fallen die Korrelationen verfälscht aus. Die Evidenz in den frühen Arbeiten für die Existenz eines Halo-Effekts war zunächst nur das Auftreten unplausibel hoher Korrelationen zwischen Eigenschaftsratings.

Einen Schritt weiter ging Newcomb (1931) und neuerdings D'Andrade (s.o.), der den Einfluß eines Halo-Effektes aus dem Unterschied zwischen Korrelationen schloß. Die eine Korrelation bezog sich auf Einstufungen von zwei Merkmalen unmittelbar nach der Verhaltensbeobachtung, die andere auf Einstufungen aus dem Gedächtnis, und diese fiel durchgängig höher aus. Was auch immer die Verfälschung bewirkte, es brauchte eine gewisse Zeit zu seiner Wirkung oder es wurde durch den unmittelbaren Bezug zum beobachteten Verhalten an seiner Wirkung gehindert.

Daß Beurteiler, Personen überhaupt, relativ stabile und interindividuell vergleichbare Vorstellungen darüber haben, wie (Persönlichkeit-)Eigenschaften kovariieren, ist spätestens seit Bruner & Tagiuri (1954) und Cronbach (1955) Ergebnis und Gegenstand der Forschung zur „impliziten Persönlichkeitstheorie“. Relevant sind solche Befunde auch für Arbeiten im Bereich der Persönlichkeitsforschung, die ihre Aussagen stark auf Korrelationen zwischen Ver-

haltenseinschätzungen stützen (z.B. Cattell 1957, Lorr & McNair, 1965). Ihr Anspruch, relativ stabile Persönlichkeitsdimensionen gefunden zu haben, wurde durch andere Forscher (z.B. Mulaik 1964, Passini & Norman 1966) in Frage gestellt, die diese Dimensionen als „Produkt stereotyper und semantischer Annahmen der Beurteiler“ deuteten (so Berman & Kenny 1976). Eine vergleichbare Problematik ergibt sich bei Korrelationen zwischen klinischen, diagnostischen Urteilen (z.B. Chapman & Chapman 1967, 1969; Starr & Katkin 1969).

Berman & Kenny (1976) gehen wieder einen Schritt weiter: Während in den früheren Untersuchungen die „wahre“ Korrelation zwischen den Merkmalen unbekannt war, variieren die Autoren diese Korrelation experimentell, indem sie ihren Beurteilern Bilder von Personen zeigten, und mit jedem Bild eine Aussage darboten, die die Ausprägung des Merkmals angab (z.B.: John ist sehr freundlich). Da John auch andere Merkmale zugeschrieben wurden, konnten Berman & Kenny den „wahren“, d.h. von ihnen hergestellten Zusammenhang zwischen den Eigenschaften variieren. Sie wählten nun nicht beliebige Merkmale, sondern solche, die nach Einstufungen anderer Vpn in einer Vorstudie paarweise stark positiv, stark negativ oder gar nicht korrelierten. Die Vpn der Hauptstudie stuften aus ihrer Erinnerung die ohne Aussagen erneut dargebotenen Personenbilder ein. Verglichen wurden die aus diesen Einstufungen berechneten Korrelationen mit den von den Autoren hergestellten.

Gleichgültig, wie hoch die experimentell hergestellte Korrelation war: Jene Merkmalspaare, die nach Meinung der Vpn der Vorstudie hoch korrelieren, zeigten auch die aus den Einstufungen in der Hauptstudie berechneten höchsten Korrelationen. Analoges galt für die negativen und Null-Korrelationen, und dies galt auch für über alle Beurteiler gemittelte Einstufungen. Als paralleler Befund wird auf die Arbeiten von Lay & Jackson (1969) und Stricker et al. (1974) verwiesen, in denen große Ähnlichkeit zwischen den Korrelationsmustern der Antworten in Persönlichkeits-Inventaren wie dem MMPI einerseits und Annahmen von Vpn über den Zusammenhang der durch diese Inventare erfaßten Merkmale gezeitigt wurde.

Die Arbeit von Berman & Kenny hat eine Diskussion ausgelöst, in die Block (1977) folgende Argumente einbringt:

- (1) Die experimentelle Manipulation der Korrelation zwischen den Merkmalen sei „künstlich“, denn die paarweisen Merkmale seien bei hoher vermuteter Korrelation Homonyme, bei negativen Korrelationen aber Gegensätze gewesen. Wie man aus Arbeiten wie denen von Cohen & Schümer (1968), Cohen (1971), Schümer (1971) entnehmen könne, bemerken Vpn meistens, wenn und daß sie widersprüchliche Information über eine zu

beurteilende Person angeboten bekommen, und das führe für zumindest einige Vpn zu einem „different context of understanding“.

- (2) Die Situation sei für die Vpn von Berman & Kenny nicht die einer Verhaltensbeurteilung gewesen, da ihnen kein Verhalten gezeigt wurde; es sei eher paired associate Lernen verlangt worden, was andere Prozesse auslöse.
- (3) Die Informationsbasis der Vpn von Berman & Kenny entspreche wegen ihrer kurzen zeitlichen Dauer allenfalls dem Bilden eines ersten Eindrucks.
- (4) Block listet Anforderungen an Einstufungen von Verhalten auf, die seines Erachtens allgemein zu erfüllen seien, bei Berman & Kenny jedoch teilweise vernachlässigt worden wären, und zwar:
 - (a) mehrere Beobachter schätzen unabhängig voneinander ein;
 - (b) Beobachter sollten ihre Urteile auf extensive, und - wegen der Wirkung von Kontexteffekten - auf verschiedenartige Gelegenheiten stützen können;
 - (c) die Beobachter sollten eine gemeinsame „evaluative metric“ benutzen, so daß es erlaubt ist, ein aggregiertes Urteil zu bilden;
 - (d) ipsative Urteile seien außennormorientierten vorzuziehen;
 - (e) „Prior calibration of the observers“, also Schulung, und konstante Kontrolle der Leistung der Beobachter sollte gewährleistet sein.

Die Erwiderung findet sich in Berman & Kenny (1977). Sie kommentieren das Argument, ihren Vpn sei kein beobachtbares Verhalten als Urteilsbasis angeboten worden, mit dem Hinweis, seit langem - so z.B. schon Guilford (1954) - hätten Psychometriker als wichtige Möglichkeit, Urteilsfehler zu reduzieren, vorgeschlagen, die Anforderungen an das Urteilsverhalten der Beobachter zu minimieren. Das sei in ihrer Studie geschehen, und trotzdem seien die ausgeprägten Verzerrungen festgestellt worden. - Zum Gedanken, Urteilsfehler durch Wiederholungen „auszumitteln“, argumentieren Berman & Kenny (1977): Um Urteile über eine Mittelwertsbildung gültiger zu machen, um also das Gesetz der großen Zahl anwenden zu können, benötige man Messungen, deren Fehler unkorreliert seien. Gerade dies sei bei Verhaltenseinstufungen in der Regel nicht der Fall. Es sei - noch anspruchsvoller - sogar zu fordern, daß die Fehler bei der Beurteilung des einen Merkmals durch den einen Beobachter unabhängig seien von den Fehlern bei den Einschätzungen eines anderen Merkmals durch einen anderen Beobachter. Analysen von Berman & Kenny legten den Schluß nahe, daß auch diese Bedingung nicht immer erfüllt sei.

Damit haben wir die Diskussion der Abhilfeschläge eingeleitet, die meist mit der Mahnung beginnen, die Urteilsskalen, ihre Etikettierung und die zugehörige Instruktion sorgfältig zu wählen. Jeder Beurteiler solle alle Personen, zunächst oder ausschließlich, nur hinsichtlich eines Merkmals einstufen. Johnson (1965) konnte jedoch auch bei diesem Vorgehen die Verzerrung nicht

vermeiden. Statistische Adjustierungen wurden von Chi (1937) und Stanley (1961) vorgeschlagen, wozu Berman & Kenny (1976) bemerken, die Korrektur geschehe bei diesen Verfahren aufgrund von Unterschieden zwischen den Beobachtern, deren „common stereotypes“ würden also nicht berücksichtigt.

Die in diesem Abschnitt behandelten Arbeiten beziehen sich auf Beurteilungen, nicht auf Wahrnehmungen, aber jede Beobachtung muß in irgendeiner Form - und in der Psychologie ist es oft die sprachliche - ihr Ergebnis mitteilen. Die Studien zeigen: Wird das Verhalten nicht sogleich (vom Beobachter) aufgezeichnet, so können systematische Verzerrungen eintreten. Erklärt wurden diese „Fehler“ zunächst als bedingt durch einen Gesamteindruck, den der Beobachter vom Beobachtungsgegenstand, einer Person, gewonnen hat, dann als bedingt durch eine implizite Theorie über das Zusammen-vorkommen der Merkmale, wobei diese die Beurteilungen steuernde „Theorie“ weder spezifisch auf einen bestimmten Beobachtungsgegenstand bezogen ist noch als charakteristisch für einen bestimmten Beobachter angesehen wird. Der Halo-Effekt wurde vom Beobachtungsfehler zu einem von vielen Indizien dafür, wie konkrete einzelne Beobachtungen mit früher gesammelten Erfahrungen integriert werden.

3.2.3 *Die Theorie der Signalentdeckung: Der Beobachter als Sensorium und als Entscheidungsinstanz*

Wir geben hier keine Darstellung auch nur der Grundzüge der theory of signal detection (erstmalig Peterson et al. 1954, van Meter & Middleton 1954), in die es mehrere gute Einführungen, z.B. Coombs et al. (1970), McNicol (1972), systematische Darstellungen (Egan 1975), und übersichten ihrer Anwendung in der Psychologie gibt (Swets 1964, Price 1966, Pastore & Scheirer 1974). Wir werden vielmehr an einem fiktiven und in einigen Punkten absichtlich unzulänglichen Beispiel die Leistungen und Möglichkeiten der Theorie der Signalentdeckung skizzieren.

Ausgangspunkt der Theorie, und gegenüber der bisherigen Psychophysik das wesentlich Neue ist die analytische Trennung zwischen dem Beobachter als einem *Sensorium*, einerseits, d.h. als Registriersystem, das mit einer gewissen, empirisch festzustellenden Sensitivität auf die externe Reizvariation reagiert, und andererseits dem Beobachter als *Entscheidungsinstanz*, d.h. als Bewertungssystem, das für seine Urteile über eine Reizgegebenheit nicht nur den sensorischen input berücksichtigt, sondern auch seine subjektiven Erwartungen und Werte. Die Erwartungen können sich auf die Wahrscheinlichkeiten beziehen, daß ein Reiz, Signal, oder nur noise, kein Signal auftritt. Die Bewertungen beziehen sich auf die belohnenden Folgen eines richtigen, und die bestrafenden Folgen eines falschen Urteils. Die Theorie wird angewendet,

wenn das Signal schwach ist relativ zu den (störenden) Hintergrundreizen. Für starke Reize, die stets eindeutig vom Hintergrund oder im Vergleich zu anderen Reizen unterschieden werden, ist die Theorie nicht erforderlich. Die Theorie legt fest, welches Urteil über den Reiz man in einer unsicheren Situation fällen sollte, indem sie Kriterien und Eingangsgrößen spezifiziert, nach denen die Wahl zwischen Reaktionsmöglichkeiten zu treffen ist.

Das Beispiel: Ein Beobachter soll immer dann eine bestimmte Kodierungsmarke benutzen, wenn eine Vp einer anderen „freundlich zulächelt“. Das freundliche Zulächeln ist in dem Sinne ein „schwaches Signal“, als es zum einen durch rein äußerliche Umstände - wenn etwa sich die Vp vom Beobachter mehr oder weniger abwendet - nur teilweise oder verzerrt als Reiz angeboten wird, zum anderen nicht präzise definiert ist, was „freundlich zulächeln“ bedeutet - es gibt keinen eindeutigen Vergleichsstandard. Andererseits ist „noise“ stets vorhanden, auch für den Beobachter zeigt die Vp ununterbrochen Ausdrucksverhalten, das bisweilen dem Zulächeln ähnlich ist und damit verwechselt werden kann. Um die Theorie der Signalentdeckung anwenden zu können, muß nun eindeutig geklärt sein, wann in der Umwelt des Beobachters der Reiz auftritt, und wann nicht. Wir stoßen hier auf ein Problem, das die Übertragung der Signalerkennungs-Theorie auf Beobachtungssituationen außerhalb des psychologischen Laboratoriums erschwert hat, und das - in anderem Zusammenhang - D'Andrade (1974, S. 159) so formuliert:

„Attempting to specify what aggressive means by defining the term with reference to more specific acts, such as hitting or insulting, still fails to make the decision process explicit, since the rules for coding behavior remain dependent on a set of undefined terms. Unless the measurement process is based solely on physical properties, at some point such verbal coding rules always fall back on undefined terms.“

Die Theorie nimmt nun weiter an, die sensorische Evidenz, auf die sich der Beobachter für sein Urteil stützt, ließe sich als ein Kontinuum zunehmender oder abnehmender Stärke darstellen, als eine „evidence variable“ (Ingleby 1974) oder „decision axis“, deren inhaltliche Zusammensetzung zunächst nicht geklärt werden muß. Über diesem Kontinuum existieren zwei Verteilungen, die eine für die Fälle, in denen das Signal vorhanden ist, die andere für die Fälle mit ausschließlich Störhintergrund. Diese beiden Wahrscheinlichkeits-Verteilungen, daß bei einer bestimmten Evidenzstärke das Signal vorhanden ist oder nicht, hat der Beobachter aus seiner Alltagserfahrung gelernt. Die Beobachtungssituation legt auch fest, sei es indirekt etwa über eine Instruktion, frühere Erfahrungen etc., sei es direkt durch experimentelle Manipulation einer Auszahlungsmatrix, wie großen Wert der Beobachter darauf legt, zutreffend „Lächeln“ und „Nicht-Lächeln“ zu identifizieren, und wie gravierend für ihn die Fehler des false alarm (er registriert ein Lächeln, das nicht auftrat) und des Übersehens sind.

Die Theorie definiert nun zwei statistisch voneinander unabhängige Kennwerte, von denen d' die Empfindlichkeit des Beobachters (hier: seine Sensitivität für zutreffende Identifikation von Lächeln aus dem übrigen Ausdrucksgeschehen) und β einen Cut-Off-Punkt auf der Entscheidungsachse darstellt, der vorschreibt, bis zu welcher Evidenzausprägung das Urteil „ich habe Lächeln beobachtet“ abzugeben sei: Die analytische Trennung zwischen dem Beobachter als Sensorium und als Entscheidungsinstanz geschieht so, daß Erwartungen und Werte des Beobachters ausschließlich in die Bestimmung von β , nicht aber von d' eingehen. Um β bestimmen zu können, muß man noch festlegen, welcher Entscheidungsregel der Beobachter als *bester* folgen soll, und dafür gibt es mehrere Möglichkeiten, etwa die, den maximal erwarteten Nutzen zu erstreben, oder den minimal zu erwartenden Verlust. Ein Beobachter, der in seinem Beurteilungsverhalten einer solchen Regel folgt, heißt rationaler *Beobachter*. Wenn er nun noch Leistungen zeigt, die nur durch die Struktur des Reizangebotes in seiner Umwelt, nicht aber durch seine Schwächen als Sensorium begrenzt werden, bezeichnet man ihn als *idealen Beobachter* (oder Empfänger). Ändert sich die Umwelt, so ändert sich die Struktur des idealen Beobachters in solcher Weise, daß auch unter den neuen Bedingungen die bestmöglichen, nicht unbedingt 100% zutreffenden Urteile über die Reizgegebenheiten gefällt werden (Egan 1975).

Empirische Untersuchungen (z.B. Coombs et al. 1970) zeigen, daß reale Beobachter deutlich schlechtere Leistungen zeigen als ideale, und daß dies u.a. wahrscheinlich an unzulänglichen Speicherungen des Signals im Gedächtnis liegt. Wir können hier nicht die ständig wachsende Zahl empirischer Arbeiten wie etwa die von Ingleby (1974) referieren; uns kam es lediglich darauf an, die Perspektive der Theorie der Signalentdeckung aufzuzeigen und ihre Anwendung auch auf nicht streng oder ausschließlich physikalisch definierte Reize zu diskutieren und anzuregen.

3.2.4 Verhaltenseinschätzungen als Testscores

Man kann, wie dies van der Kamp & Mellenbergh (1976) tun, Verhaltenseinstufungen als „a special class of test scores“ auffassen, und dann konsequent annehmen „... that each rater may be regarded as a test instrument“ (S. 311). Mit der Anwendung der Testtheorie (s. Kristof und Fischer in diesem Band) betrachtet man dann die Frage, in welcher Hinsicht Beobachter übereinstimmen. Jeder Beobachter wird dabei zu einer Methode, und gefragt ist nach der Äquivalenz verschiedener Methoden, eine „zugrundeliegende“ oder „hypothetische“ Variable zu messen.

Die strengste Form der Äquivalenz ist die Austauschbarkeit (ratings are interchangeable), die sich mit einem Kriterium nach Wilk-Votaw (s. Gulliksen 1950, 1968) prüfen läßt. Für die Urteile von Beobachtern bedeutet dies, sie

seien in der gleichen Weise - am gleichen Skalenursprung - verankert (gleiche Mittelwerte), benutzten die gleiche Skaleneinheit (gleiche Varianzen), erfaßten die Beziehungsstruktur zwischen verschiedenen Variablen in der gleichen Weise (gleiche Kovarianzen), und ihre Einstufungen seien deshalb auch für alle Beobachter gleich verläßlich und gleichermaßen gültig. Diese unrealistischen Anforderungen werden nun von den Autoren Schritt für Schritt abgeschwächt, unterschiedliche Reliabilität der Beobachter wird eingeräumt, etc.

Werts et al. (1976) gehen ebenfalls von der klassischen Testtheorie aus und betrachten den Fall, daß mehrere Beobachter die gleichen Objekte in mehr als einer Hinsicht einschätzen. Unter diesen Umständen sei es für einen Beobachter kaum möglich, seine Einschätzungen auf der einen Dimension nicht von seinen Kenntnissen seiner Urteile auf den anderen Dimensionen beeinflusst sein zu lassen. „This kind of contamination means that the errors of measurement for one dimension may be correlated with the errors on other dimensions, i.e., the *intrajudge* measurement errors are correlated.“ (S. 319). Dann sind die Kovarianzen zwischen Einstufungen auf verschiedenen Dimensionen durch den gleichen Beobachter nicht gleich den Kovarianzen zwischen den zugrunde liegenden wahren Werten, wie normalerweise in der klassischen Testtheorie angenommen werde. Auch die übliche Minderungskorrektur um die Korrelation zwischen wahren Werten verschiedener Dimensionen zu bestimmen, sei nicht anwendbar. Die Autoren entwickeln ein Analyseverfahren, in dem - unter der Annahme, die Fehler eines einzelnen Beobachters seien korreliert, nicht jedoch die zwischen verschiedenen Beobachtern - sich die Korrelationen zwischen den wahren Werten auf verschiedenen Dimensionen, die Reliabilitäten jedes Beobachters auf jeder Dimension und die Korrelationen zwischen den Urteilsfehlern des einzelnen Beobachters berechnen lassen. Es interessiert in diesem Ansatz also nicht die psychologische Ursache für meist wohl überhöhte Korrelationen zwischen verschiedenen Variablen, die am gleichen Beobachtungsgegenstand erfaßt werden; vielmehr wird gezeigt, unter welchen Annahmen über den Beobachter man dennoch die „wahren“ Zusammenhänge bestimmen könne.

3.2.5 Brunswiks probabilistischer Funktionalismus: Beobachtung als Leistung

Als ein Beispiel für die Ansätze der kognitiven Psychologie, die das Zustandekommen von Wahrnehmungsurteilen detailliert beschreiben, erwähnen wir kurz Brunswiks Modell (Brunswik 1952, 1955, 1956; als gute Einführung: Postman & Tolman 1959; wir folgen weitgehend dieser Darstellung). Brunswik unterscheidet die einer zentralen und motorischen Reaktion vorhergehenden Bedingungen als Reize mit unterschiedlicher Entfernung vom Organismus. Er betrachtet zum einen distale Reize, Gegebenheiten der Umwelt, mit denen der Organismus keinen unmittelbaren Kontakt hat, zum anderen proxi-

male Reize, die an den Grenzen des Organismus, beispielsweise als Netzhautbild gegeben sind.

Brunswik untersucht nun drei funktionale Beziehungen. Als erste sei die zwischen distalen und proximalen Reizen betrachtet, die dann als cues und als proximale Wirkungen erscheinen. Einige Hinweisreize variieren mit größerer Wahrscheinlichkeit systematisch als Funktion von distalen Reizänderungen. Die Stärke der Kovariation zwischen proximalem Hinweisreiz und distalem Merkmal definiert die ökologische Validität dieses Hinweisreizes hinsichtlich jenes Merkmals. Die ökologischen Validitäten beschreiben die ererbte und erlernte - u. U. auch im Beobachtungstraining - Einbettung des Beobachters in seine Umwelt. Sie zu kennen ist für Schlüsse aus Beobachtungen auf das Beobachtete wesentlich. Die zweite Funktion beschreibt als Beziehung zwischen proximalen Wirkungen und zentralen, perzeptiven Reaktionen den Gebrauch, den der Beobachter von Hinweisreizen macht. Um sich an distale Gegebenheiten anzupassen, gebraucht der Organismus die ihm verfügbaren proximalen Hinweisreize. Da diese nur begrenzte ökologische Validität haben, muß der Organismus - und, vielleicht bewußter, der Beobachter - eine Regel finden, wie er verschiedene Hinweisreize gewichten und kombinieren soll, um die distalen Gegebenheiten möglichst richtig abzuschätzen. Die dritte Beziehung, zwischen distalen Merkmalen und perzeptiver Reaktion, wird als funktionale Validität bezeichnet. Sie quantifiziert das Ausmaß, in dem der Organismus die Wahrnehmung der Umwelt geleistet hat.

Bei der Auseinandersetzung mit einer Umwelt kann man demnach Beobachtung als Leistung in zweifacher Hinsicht untersuchen: als Transformation von distalen in proximale Reize und als Integration verschiedener Hinweisreize zu einer kognitiven Reaktion. Um die Leistung eines Organismus einschätzen zu können, muß man ihn in seinem natürlichen Habitat untersuchen, denn dort treten die Hinweisreize in jener Kovariation auf, an die der Organismus seine Anpassung vollzogen hat. Werden die Hinweisreize, wie oft bei Wahrnehmungsexperimenten im Labor, isoliert und voneinander unabhängig variiert, dann müssen die Erkenntnisse über Teile des Systems kein zutreffendes Bild vom Gesamtsystem ergeben. Ein im Sinne Brunswiks repräsentativer Versuchs- und Beobachtungsplan sollte daher eine unverfälschte Stichprobe von Situationen aus der Ökologie des Organismus ziehen (Anwendungen der Brunswikschen Gedanken in der Lerntheorie unter dem Stichwort *multiple-cue probability learning task* oder *probabilistic concept identification task*, in der Diagnostik s. Lüer & Kluck in diesem Band; formale Weiterentwicklungen z.B. bei Castellan 1973, Steward 1976).

3.3 Die Wahl von Beobachtungseinheiten durch Beobachter

Im Artikel über die wissenschaftliche Beobachtung wurde bereits die Frage angesprochen, wie die *Beobachtungseinheit* vom Forscher zu definieren und dem Beobachter vorzugeben sei. Wir wiesen darauf hin, daß der Forscher keine vollständige Kontrolle darüber haben kann, wie der Beobachter seine Beobachtungseinheiten festlegt. Dann ist es also eine empirische Aufgabe, die tatsächlich gewählten Einheiten zu bestimmen und die Regeln zu finden, wie Einheiten zustande kommen. Mit dieser Frage, wie - in welche Einheiten - Beobachter das Beobachtete gliedern haben sich in letzter Zeit Newton und seine Mitarbeiter befaßt (Newton 1973, 1976; Newton & Engquist 1976; Newton & Rindner 1979; Newton et al. 1977). Nach Newton gliedern Beobachter den Verhaltensstrom, indem sie aufeinanderfolgende Definitionspunkte wählen. Diese points of definition, auch breakpoints genannt, sind Stellen, die die Wahrnehmende als jene identifiziert, an denen sich Handlungen ereignet haben.

Beobachter scheinen diese Punkte so auszuwählen, daß sie eine Zusammenfassung der Information darstellen, die sie aus der Beobachtung der Handlungsabfolgen gewonnen haben. Newton (1973) geht davon aus, daß Wahrnehmende aktiv die Informationsaufnahme bei der Beobachtung kontrollieren, indem sie eine größere oder kleinere Anzahl von Definitionspunkten wählen, also ihre Einheit der Segmentierung größer oder kleiner festlegen. Wie kann man nun nachweisen, daß „observational units“ existieren, und wie kann man einzelne konkrete Beobachtungseinheiten identifizieren? Wenn der Strom des beobachteten Verhaltens ununterscheidbar kontinuierlich erschiene, könnte weder der gleiche Beobachter bei wiederholten Gelegenheiten noch verschiedene Beobachter am gleichen Material übereinstimmend Markierungspunkte feststellen. Die Untersuchungen von Newton und Mitarbeitern zeigen jedoch hohe intraindividuelle Konsistenz und hohes interrater-agreement, wenn Verhalten in Einheiten unterteilt werden sollte. Die Vpn mußten Grenzen bestimmen, durch die eine „behavior unit“ von einer anderen getrennt wird (Instruktion: „... press the button whenever, in your judgment, one meaningful action ends and a different one begins“). Ferner konnte diese Forschergruppe zeigen, daß diese Definitionspunkte andere Merkmale aufwiesen als willkürlich aus der Verhaltenssequenz herausgegriffene andere Punkte. Wurden kurze Segmente aus Filmen herausgeschnitten, so konnten die Auslassungen richtiger entdeckt werden, wenn Definitionspunkte entfernt worden waren als bei anderen Punkten. Wurden Bilder verwendet, die Definitionspunkten entsprachen, so ließ sich aus ihnen das Geschehen leichter und zutreffender interpretieren als aus anderen Bildern, und die Reihenfolge der Bilder ließ sich für Definitionspunkte eher richtig bestimmen. Definitionspunkt-Bilder werden auch besser im Gedächtnistest wiedererkannt. Diese Punkte, so schließen die Autoren aus diesen Befunden, stellen kein Artefakt dar, sondern stützen die

Annahme, Verhalten werde als in Einheiten gegliedert wahrgenommen und geschildert. Definitionspunkte werden als Reize mit besonders hoher Information für den Beobachter aufgefaßt. Mit dem Signal, daß eine neue Einheit beginnt, ist auch Information verbunden, wie es weiter geht. Für die Antizipation oder Vorhersage des neuen Geschehens wird durch die Information im Definitions-Reiz eine Basis geschaffen, die im Verhalten der Beobachteten vor dem Auftreten dieses Punktes nicht vorhanden war. Für das Verstehen von Verhalten wäre demnach notwendig, die Information zu kennen, die durch die Definitionspunkte geboten wird. Für vertraute Verhaltenseinheiten, die in sich relativ abgeschlossen waren, weil sie z.B. die Erledigung einer einfachen Aufgabe mit ersichtlichem Zweck erforderten, konnten Newton und Mitarbeiter ihre soeben geschilderte Auffassung gut belegen. Schwierigkeiten ergaben sich für Verhaltenssequenzen mit rhythmischen Wiederholungen, z.B. Tanz zu Rockmusik.

Inhaltlich lassen sich Definitionspunkte allgemein kennzeichnen als die Stellen, an denen sich gegenüber zuvor diskriminierten Zuständen Veränderungen ergeben. Vpn, die spontan oder auf Instruktion hin kleinere Einheiten kodieren, wählen nicht andere, sondern feiner aufgelöste Einheiten. D.h., es gibt eine Hierarchie der Einheiten; bei der größten Einheitenbildung (unitization) werden kleinere zusammengefaßt, und zwar nach übergeordneten Zielen, die dem Beobachteten zugeschrieben werden, Unterziele bestimmen die kleineren Einheiten. Kleine Einheiten werden spontan gewählt, wenn Verhalten irregulär, lose organisiert erscheint, was mit der Annahme übereinstimmt, Definitionspunkte hätten eine wichtige Funktion für die Vorhersagbarkeit. Bleibt das Verhalten gleichförmig, so werden allmählich größere, zeitlich längere Einheiten gewählt. Treten unerwartete Ereignisse ein, so kehrt der Beobachter zur Feingliederung zurück.

Wie auch Wilder (1978) stellten Newton und seine Mitarbeiter einen Trend fest, mit fortschreitender Beobachtungsdauer zunehmend größere Einheiten zu wählen. Anscheinend fühle sich der Beobachter nach einiger Zeit für seine Zwecke hinreichend informiert, und weitere Information werde dann nicht mehr beachtet, oder der Aufwand, sie zu verarbeiten, lasse nach. Die vom Beobachter gewählte Größe der Beobachtungseinheit bestimmt dann die obere Grenze, wieviel Information gewonnen und wie rasch Zufriedenheit mit dem eigenen Informationsstand erreicht werden kann.

In diesem Zusammenhang seien einige Befunde über die Beziehungen zwischen skaliertem Informationsniveau und klinischen Urteilen berichtet, die auf Beobachtungen (meist Filmen) basierten (Feger, 1972, Kap. 7). 1. Je länger die Darbietungsdauer eines Filmes über eine Person in Testsituationen, je ausgedehnter also die Möglichkeit zur Beobachtung war, als desto höher wurde die vorhanden geglaubte Informationsmenge eingestuft. Die Zunahme der Informationsmenge erwies sich als kurvilinear, negativ beschleunigt. 2. Es gibt über

verschiedene Zeitpunkte der Informationsaufnahme und über verschiedene zu beurteilende Merkmale konstante interindividuelle Unterschiede in der subjektiven Informationsmenge, über die ein Beobachter zu verfügen glaubt. 3. Beobachter, die sich besser informiert glauben, stufen Verhalten extremer ein; ein Zusammenhang mit Konfidenz - der subjektiven Sicherheit, richtig beurteilt zu haben - besteht nicht. 4. Beobachtertraining erhöht die Konfidenz.

3.4 Der Entstehungsprozeß von Beschreibungen

Alle Beobachtungen, alle Wahrnehmungen schlechthin, müssen wenigstens kurzzeitig im Gedächtnis gespeichert werden, damit sie berichtet werden können. Die typische Art, Beobachtungen zu berichten, ist die verbale Beschreibung. Ericsson & Simon (1980) fassen Beschreibungen von Beobachtungen (verbal reports) als Daten auf, d.h. sie fordern - um die Aussagefähigkeit der Beschreibungen abschätzen zu können - eine Erklärung der Prozesse, wie verbale Berichte zustande kommen, und von welchen Bedingungen sie abhängen. Wie man bei diesen Autoren erwarten darf, legen sie eine kognitive *Rahmentheorie des Entstehungsprozesses von Beschreibungen* vor, deren Grundannahmen sich folgendermaßen skizzieren lassen: Man muß wissen, was ein Beobachter im Kurzzeitgedächtnis gespeichert hat, denn das kann er *verlässlich* verbalisieren. Bezieht sich die Frage eines Forschers (oder eines Beobachters an sich selbst) jedoch auf Material, das nicht gespeichert ist, so muß er schlußfolgern, und Schlüsse dieser Art können falsch sein. Aufforderungen wie „Denken Sie laut“ ändern kognitive Prozesse nur dann, wenn durch sie eine Vp gezwungen wird, ihre Aufmerksamkeit auf Gegebenheiten zu richten, denen sie sich ohne eine solche Instruktion nicht zugewandt hätte.

Ericsson & Simon beginnen ihre Argumentation im einzelnen mit einem Hinweis auf die „schizophrene“ Behandlung verbaler Berichte durch Behavioristen. In der Forschung über Begriffsbildung beispielsweise werde als hartes Datum akzeptiert, wenn eine Vp mit ja oder nein die Frage beantwortet, ob die Vorlage eine Instanz des Konzeptes sei, aber nicht akzeptiert, wenn sie berichtet, sie prüfe die Hypothese, das gesuchte Konzept sei „ein kleiner, gelber Kreis“. Es fehle ein klares *Kriterium*, was denn nun ein akzeptabler Report sei und was nicht. Ein solches einfaches Kriterium geben nun auch Ericsson & Simon nicht. Sie wenden das Problem vielmehr so, daß ein Verwender von Berichten aus deren Entstehungsprozeß beurteilen muß, ob sie für seine Zwecke brauchbar sind. Die Frage ist dann also nicht mehr, ob eine Vp „wahrheitsgetreu“ berichtet, sondern vielmehr, wie sie dazu kommt, etwas Bestimmtes auszusagen.

Unterschiede im Entstehen von Berichten ergeben sich aus den Umständen, unter denen sie gewonnen wurden. Die Autoren klassifizieren diese Umstände

zunächst danach, ob es die hauptsächliche Aufgabe der Vp war, einen verbalen Bericht zu geben, oder ob die Beschreibung nur ein Nebenprodukt war, das bei der Bewältigung einer anderen, der eigentlichen Aufgabe anfiel. Die Verbalisierung kann weiter gleichzeitig mit der Aufgabenbewältigung geschehen oder retrospektiv. Dabei kann man der Vp gezielte Fragen stellen oder solche, die von ihr Verallgemeinerungen verlangen. Einige Einschränkungen für die Benutzung von Berichten ergeben sich sogleich aus dieser Klassifikation: Ist die Verbalisation nur Nebensache, können die Berichte unvollständig und für die Fragestellung des Forschers irrelevant sein. Steht die Verbalisation im Vordergrund, so besteht die Gefahr der Interferenz mit der Bewältigung der Versuchsaufgabe. Bittet man die Vp um generelle Aussagen, so wird sie um so mehr schlußfolgern statt berichten, je allgemeiner und umfassender ihre Aussage sein soll. Wird der Bericht retrospektiv verlangt, dann werden Vollständigkeit und Gültigkeit davon abhängen, in welchem Umfang bei der früheren Aufgabenbewältigung zufällig das den Forscher Interessierende gespeichert wurde.

Wie in der neueren Psychophysik wird unterschieden zwischen der inneren Repräsentation der gespeicherten Information und dem Bericht, und zwischen beiden findet der Prozeß der Verbalisation statt. Je nach der einer Vp gestellten Berichtsaufgabe kann dieser Prozeß bedeuten, (1) daß noch nicht verbal gespeicherte Information in verbale Form recodiert werden muß, (2) daß Such- und Filterprozesse ablaufen müssen, wenn der Bericht bestimmte Anforderungen erfüllen soll, z.B. nur dieses, nicht aber jenes, oder in einer bestimmten Form zu schildern, und (3) daß Abstraktionen und Verallgemeinerungen vom Beobachter verlangt werden.

Ericsson & Simon klassifizieren gängige Methoden, wie Forscher von Personen Berichte erfragen (s. Tab. 2). Für die Unterscheidungen und Beispiele im einzelnen muß auf die Veröffentlichung verwiesen werden. Wesentlich für die Bewertung der Verfahren - es ergeben sich um so mehr Vorbehalte, je mehr man in der Tabelle nach unten und nach rechts geht - ist das Ausmaß, in dem Vpn statt sich zu erinnern, um Urteile gebeten werden, und je weniger Vpn sich tatsächlich erinnern können, weil die erforderliche Information nicht mehr im Kurzzeitgedächtnis ist, nur unvollständig dorthin aus dem Langzeitgedächtnis zurückgeführt werden kann, oder nie im Gedächtnis war.

Das methodische Vorgehen beim Gewinnen der Berichte ist nur ein Gesichtspunkt, nach dem deren Qualität beurteilt werden kann. Ein anderer ist, die psychologischen Bedingungen zu berücksichtigen, unter denen Wahrnehmungen zustande kommen und zu Berichten werden können. Die (schon bestehenden) kognitiven Strukturen und die bei einer Beobachtungsmöglichkeit ablaufenden Prozesse - wie Schwankungen der Aufmerksamkeit, Störungen durch die Umwelt, Automation der Vollzüge - bestimmen, was bewußt werden kann, und somit auch, worauf sich die Aufmerksamkeit richten kann. Das

Tabelle 2: Eine Klassifikation verschiedener Arten von Verbalisierungsaufgaben als Funktion des Zeitpunktes der Verbalisierung (Zeilen) und der Abbildung von beachteter auf verbalisierte Information (Spalten); Tabelle 1 bei Ericsson & Simon (1980, S. 244)

Beziehung zwischen beachteter und verbalisierter Information				
Zeitpunkt der Verbalisierung	Direkt eins zu eins	Mittelbare Verarbeitung		
		viele zu eins	unklar	keine Beziehung
Aufmerksamkeit der Information zugewendet	Laut Sprechen Laut Denken			
		mittelbare schlußfolgernde und generative Prozesse		
Information befindet sich noch im Kurzzeitgedächtnis	gleichzeitiges Erfragen			
Nach Abschluß der aufgabenbezogenen Prozesse	Retrospektives Erfragen	Erbitten allgemeiner Berichte	Erfragen hypothetischer Zustände	Erfragen allgemeiner Zustände

wiederum legt fest, was wie gespeichert wird und abgerufen werden kann, wenn ein Bericht gewünscht wird. Alle Faktoren, die fördernd oder hemmend auf Speichern und Abrufen wirken, beeinflussen deshalb auch positiv oder negativ eine zutreffende Wiedergabe (kognitiver) Gedächtnisinhalte. So wird die These begründet, es gebe Bereiche, in denen im allgemeinen und besonders nach einer Einübungsphase die Prozesse so schnell und ohne Aufmerksamkeitszuwendung ablaufen, daß ein zutreffender Bericht nicht möglich sei. Das wird beispielsweise für Enkodierungsprozesse bei der Wahrnehmung, beim Retrieval aus dem Gedächtnis und bei motorischen Prozessen angenommen.

3.5 Verhaltenseinschätzung (behavioral assessment)

Im folgenden schildern wir aus methodischer Sicht einige Ergebnisse der Studien über Verhaltensbeobachtung, die im Zusammenhang mit verhaltensthera-

peutischen Zielen durchgeführt wurden (Überblick Ciminero 1977). Wir möchten aus dieser Forschung jene Arbeiten zusammenfassen, die Einflüsse auf Beobachtungen berichten, welche sich aus der Beobachtungssituation, aus dem spezifischen beobachteten Verhalten und aus Merkmalen der Beobachter ergeben. Verhaltensbeobachtung ist für Auswahl, Steuerung und Bewertung der Therapie erforderlich. Nach dem Selbstverständnis der Forscher in diesem Bereich unterscheidet sich ihr Vorgehen von der traditionellen Persönlichkeitsforschung und der dort üblichen Verhaltensbeobachtung u.a. dadurch, daß sie nicht versuchen, zugrundeliegende Einheiten, Faktoren, der Persönlichkeitsstruktur aus dem Verhalten zu erschließen, vielmehr versuchen, die Bedingungen zu erfassen, die in spezifischen Situationen ganz bestimmte Verhaltensweisen auslösen (ausführlich in Gottfried & Kent, 1972). Peterson (1968, S. 114) gibt eine konzise Zusammenfassung:

„The central features of the method are (1) systematic observation of the problem behavior to obtain a response frequency baseline, (2) systematic observation of the stimulus conditions following and/or preceding the behavior, with special concern for antecedent discriminative cues and consequent reinforcers, (3) experimental manipulation of a condition which seems functionally, hence causally, related to the problem behavior, and (4) further observation to record any changes in behavior which may occur.“

Charakteristisch ist also, daß dem Anspruch nach der Bereich dessen, was beobachtet wird, hier wesentlich erweitert ist: die Ereignisse in der Umwelt, die dem gezeigten Verhaltenssegment vorausgehen und nachfolgen, werden ausdrücklich (und nicht nur in der Deutung durch den Beobachter implizit) hinzugenommen, und man greift in diese Umwelt gezielt ein, um die Gegebenheiten herauszufinden, die das Verhalten beeinflussen.

Zu den Methoden der Verhaltenseinschätzung werden außer den hier nicht besprochenen physiologischen Messungen Berichte der untersuchten Person (self report, self recording, hier Selbstberichte genannt) sowohl des eigenen Erlebens als auch insbesondere des eigenen Verhaltens und die Fremdbeobachtungen, gewöhnlich durch den Therapeuten, gerechnet. Zu den *Selbstberichten* zählen Interviews des Patienten durch den Therapeuten über das Verhalten des Patienten (behavioral interviews), die Anwendung von Verhalten-Fragebögen (Surveys and inventories) und die Anwendung von Registrierverfahren durch den Patienten (self-monitoring procedures), womit aus ethischen oder praktischen Gründen nicht direkt beobachtbare Verhaltensweisen, aber auch Erleben - wie der Wunsch, sich eine Zigarette anzuzünden - erfaßt werden sollen. *Fremdbeobachtungen* werden danach unterschieden, ob sie in der natürlichen, alltäglichen Umgebung (naturalistic setting) der untersuchten Person angestellt werden oder im Labor, und zwar dort in Situationen, die dem Alltag des Patienten nachgebildet sind (contrived analogue settings). Wir behandeln drei methodische Themenbereiche, die für alle Beobachtungsarten

relevant sind und vergleichsweise ausführlich untersucht wurden, und zwar Fragen der Reliabilität, der Reaktivität und der Auswirkung von Beobachtererwartungen.

3.5.1 Die Verlässlichkeit von Selbstberichten und Fremdbeobachtungen

Da Beobachter meist für die Erfassung einer *bestimmten Art von* Verhaltensweisen geschult werden, und ein häufiges Erfolgskriterium der Schulung die Reliabilität ist, stellt sich die Frage, ob eine einmal erreichte Verlässlichkeit erhalten bleibt, wenn sich - wie etwa bei Schizophrenie und manisch-depressiven Störungen - das Verhalten der Beobachteten deutlich ändert. Redfield & Paul (1976) berichten gleichbleibende Reliabilität. Kent & Foster (1977) berichten, in einer neuen Umgebung fiel die Übereinstimmung zwischen Beobachtern zunächst ab und stieg dann wieder an. Verschiedene Verhaltensarten, so wird in der Literatur häufig vermutet, jedoch selten geprüft, sollten sich unterschiedlich reliabel beobachten lassen. Jones et al. (1975) nehmen an, die Verlässlichkeit sei eine Funktion der Komplexität des Verhaltens, wobei sie Komplexität durch das vorgegebene Kategoriensystem als Anzahl der Unterscheidungen definieren, die ein Beobachter bei der Datenaufnahme treffen muß. Jones und Mitarbeiter fanden höhere Komplexitätswerte in Sitzungen kurz vor und nach Sitzungen mit Reliabilitätsprüfung. Die Auswirkung der „Komplexität“ des Kategoriensystems untersuchten Mash & McElwee (1974), indem sie verbales Material nach zwei jeweils erschöpfenden Kategoriensystemen beurteilen ließen. Das erste enthielt nur vier Kategorien, beim zweiten waren diese Kategorien noch jeweils unterteilt in zwei weitere. Höhere Übereinstimmung fand sich für das System mit weniger Kategorien. Schulung an einem kurzen Videotape, das gut vorhersagbares Interaktionsverhalten zeigte, führte zu einer geringeren Übereinstimmung bei der Beurteilung eines später gezeigten Bandes als Training an einem Tape mit wenig vorhersagbarer Interaktion (bestätigt in Mash & Makohoniuk 1975).

Gut belegt ist der positive Einfluß der Erwartung der Beobachter, hinsichtlich der Reliabilität kontrolliert zu werden, auf eben diese Reliabilität (Reid 1970, Romanczyk et al. 1973). Taplin & Reid (1973) ließen 18 undergraduate Studenten die Interaktion zwischen Mutter und Kindern auf Videotape beobachten. Nach einer Vorschulung wurden sie in drei Gruppen eingeteilt, die sich durch folgende Informationen unterschieden: I Die Versuchsleiter würden die Reliabilität nicht prüfen. II Jeder beliebige Teil des Protokolls könne später vom VI geprüft werden. III Es gäbe Stichproben (spot checks), die zuvor als solche angekündigt würden. Unter Bedingung II zeigte sich die höchste Übereinstimmung, ein mittleres Maß unter I und geringstes unter III, außer für jene Protokollteile, für die eine Kontrolle zuvor angekündigt wurde. Taplin & Reid (1973, S. 553) schlußfolgern:

„The major thrust of the findings . . . is to point to the danger of assuming that unmonitored observational data are reliable. Instead of being established a priori, the reliability of observational data ought to be empirically demonstrated. Without such demonstrations, the accuracy of observational data can only be a matter of faith.“

Beobachter können die Interpretation eines Kodiersystems zum einen während der Schulung auch im direkten Vergleich ihres Verhaltens mit dem anderen Beobachter lernen. Zum anderen können sie aber auch während der Anwendungsphase indirekt lernen, etwa über mitgeteilte Reliabilitätskoeffizienten oder inter rater agreement. Offensichtlich passen sich länger zusammenarbeitende Beobachter untereinander im Gebrauch von Kodiersystemen an. Dafür sprechen sowohl Befunde, nach denen im Verlauf längerer Beobachtungen die Übereinstimmung in einem Beobachterteam anwächst als auch der Befund, daß die Übereinstimmung in einem Team meistens größer ist als zwischen Beobachtern aus verschiedenen Teams, selbst wenn alle Beobachter zuvor durch die gleiche Schulung gegangen sind (vgl. Kent & Foster 1977).

Wir wenden uns nun der *Reliabilität von Selbstberichten* zu. Eine typische Aufgabe beim self recording besteht beispielsweise für einen Patienten darin zu registrieren, wie oft er sein eigenes Gesicht berührt. Die Reliabilität, definiert als Übereinstimmung mit einem externen Beobachter, ist in der Regel dann relativ hoch ($r = .8$), wenn die Patienten wußten, daß ihre Zuverlässigkeit geprüft wurde, und wenn ihre Zuverlässigkeit immer wieder belohnt wurde (Fixen et al. 1972, Lipinski & Nelson 1974, Lipinski et al. 1975, Nelson et al. 1975, Taplin & Reid 1973). Vergleicht man die Reliabilität von Selbstberichten mit der von Daten externer Beobachtungen, so zeigen sich Selbstberichte oft als weniger zuverlässig. Simkins (1971) erwägt dafür folgende Gründe: Externe Beobachter erhalten ein besseres oder anderes Training als Selbstberichter, sie stehen unter verschiedenartigen Belohnungskontingenzen und benutzen verschiedene Kriterien für Beurteilung und Registrierung. Bei Selbstberichtern könnten Reaktionen eine Registrierung behindern, die es für externe Beobachter nicht gibt, beispielsweise könnte das zu registrierende Interaktionsverhalten die Teilnehmer, nicht aber externe Beobachter ablenken (s. Cavior & Marabotto 1976).

Für eine *Zusammenfassung* zahlreicher, auch hier nicht berichteter Befunde erscheint es dem gegenwärtigen Methodenstand in diesem Bereich angemessen, wenn man bei der Diskussion der Reliabilitätsfrage davon ausgeht, daß ein Training die Beobachter schon bis zu einem Minimum an Verlässlichkeit führt, und dann nach Bedingungen gesucht wird, welche diesen einmal erreichten Stand beeinflussen können. Die Befunde sprechen dafür, daß Faktoren, die zwischen Schulung und Einsatz liegen, und solche, hinsichtlich derer sich Situationen der Schulung von denen des Einsatzes unterscheiden, deutliche Unterschiede in der Reliabilität bewirken. Zu diesen Faktoren gehören insbe-

sondere Erwartungen der Beobachter, die deshalb später in einem eigenen Abschnitt besprochen werden.

3.5.2 *Reaktivität*

Wir definieren Reaktivität (bewußt quantitativ) als das Ausmaß, in dem ein Verfahren das Phänomen ändert, das mit ihm untersucht wird. Wir schildern hier nicht allgemein die umfangreiche Literatur zu Versuchsleitereffekten (z.B. Rosenthal 1966, 1976; Rosenthal & Rosnow 1969) oder zu nichtreaktiven Meßverfahren (Webb et al. 1966), obwohl vermutlich vieles, das sich in dieser Literatur auf strikt experimentelles Vorgehen bezieht, auch für Beobachtungsstudien relevant sein kann. Vielmehr beschränken wir uns hier auf Arbeiten, die Reaktivität speziell bei Beobachtungen zum Zweck der Verhaltenseinschätzung untersucht.

Zunächst befassen wir uns mit der möglichen Reaktivität von Beobachtungen, die durch *externe Beobachter* angestellt werden. Nach Kent & Foster (1977) können folgende Faktoren der Reaktivität entgegenwirken: (1) Gewöhnung an die Beobachter, dies wurde allerdings in einer Studie gezeigt, in der die Gewöhnung über mehrere Wochen hinweg möglich war, bevor die Beobachtung selbst begann; (2) ein Beobachtungsplan mit häufigen und nicht vorhersagbaren Beobachtungsphasen; (3) Verzögerung des feedbacks über Beobachtungsergebnisse an die Beobachteten. Auch unter sonst gleichen Beobachtungsbedingungen und an den gleichen beobachteten Personen kann man offensichtlich nicht bei allen *Variablen* in gleichem Ausmaß Reaktivität erwarten (Roberts & Renzagha 1965).

Beim typisch experimentellen Vorgehen interagieren in der Regel nur Versuchsperson und Experimentator, der oft mit dem Forscher identisch ist. Wenn jedoch Beobachter und Forscher nicht identisch sind, etwa wenn Eltern über ihre Problemkinder berichten sollen, kann sich Reaktivität nicht nur im Verhalten der Beobachteten zeigen und i. w. S. auch beim Forscher, sondern auch beim Beobachter. Johnson & Lobitz (1974) analysieren diese Beobachtungssituation mit Hilfe des Konzeptes der *demand characteristic*: Welche spezifischen Situationsmerkmale ergeben sich für alle Personen in einer Situation mit einem Beobachter aus dessen Anwesenheit? Johnson & Lobitz nehmen beispielsweise für Eltern an, die ihr Kind als therapiebedürftig beschrieben haben, für sie ergäbe sich bei der Beobachtung in der Familie vielleicht unbewußt der Wunsch, ihr Kind tatsächlich als problematisch erscheinen zu lassen. Die Autoren baten Eltern in Familien ohne Problemkinder, ihre Kinder entweder als „brav“ oder als „ungezogen“ erscheinen zu lassen. Johnson & Lobitz konnten entsprechende Unterschiede im Verhalten von Eltern und Kindern nachweisen und argumentieren, wenn es unterschiedliche demand

characteristics gäbe, seien Personen in der Lage, darauf unterschiedlich zu reagieren und somit die Validität von Beobachtungen zu gefährden.

Im letzten Jahrzehnt hat man sich besonders um Techniken, aber auch um Versuchspläne bemüht, die eindeutige Schlüsse über Reaktivität zulassen. Die Verwendung von Kleinst(radio)sendern wurde mehrfach untersucht, z.B. von Moos (1968) an psychiatrischen Patienten. Moos verglich das Verhalten bei eingeschaltetem und bei ausgeschaltetem Sender. Insgesamt zeigte sich wenig Reaktivität, ihr Ausmaß war jedoch abhängig von der jeweiligen Situation und zeigte sich stärker bei den stärker gestörten Patienten. Johnson & Bolstad (1975) schnitten die Interaktion von Familienmitgliedern auf Tonband mit und fanden keine Unterschiede bei Anwesenheit oder Abwesenheit eines externen Beobachters. Goldfried & Linehan (1977) wenden in ihrem hier zugrundegelegten Überblick gegen beide Studien das Fehlen einer Kontrollgruppe ein, die eindeutig nicht beobachtet wurde, zumindest nach Meinung der betroffenen Vpn. In der Studie von Moos wußten die Vpn, daß sie auch dann beobachtet wurden, wenn sie den Sender nicht trugen, und in der Untersuchung von Johnson & Bolstad lief das Tonband auch dann, wenn kein Beobachter anwesend war. Diesen Einwand berücksichtigen - vielleicht ethisch bedenklich - Hagen et al. (1975). Sie arbeiteten mit einem verborgenen Mikrophon, das ständig die Interaktion von Personal und Patienten aufnahm, und verglichen das Verhalten bei An- und Abwesenheit eines Beobachters, ohne Unterschiede zu finden. Diesen Versuchsplan entwickelten Johnson et al. (1976) weiter, die Kinder mit einem Sender versahen, der das verbale Verhalten auf Tonband übertrug. Das Tonband wurde nach einem time sampling Schema ein- und ausgeschaltet, das weder dem Kind noch den übrigen Familienmitgliedern bekannt war.

Während die Daten externer Beobachter beim behavior assessment kaum durch Reaktivität beeinflusst zu werden scheinen, sind Reaktivitätseffekte bei *Selbstberichten* so stark - sie reduzieren meistens die problematische Verhaltensweise - daß self monitoring inzwischen regelmäßig, wie Goldfried & Linehan berichten, als therapeutische Technik eingesetzt wird. Erklärt wird dies damit, daß der Selbstbericht zu einem Spezialfall von feedback wird, das das Verhalten ändert. Statt auf feedback legen Ciminero et al. (1977, p. 208) mehr Wert auf die Tatsache, daß „... the presence of an observer alters the usual stimulus situation, thereby producing behavior changes“ und „... when an individual begins to self-observe his own behavior, there is also a change in the usual stimulus situation.“ Es wäre interessant zu erfahren, worin genau die Veränderung besteht, und warum sie in beiden Situationstypen so verschieden ist, daß so starke Unterschiede in der Reaktivität auftreten.

Ciminero et al. geben einen Überblick über verschiedene *Versuchspläne*, die zur Analyse von Reaktivität bei Selbstberichten verwendet wurden. Als Beispiel für eine *Einzelfallstudie* sei die Arbeit von Maletzky (1974) erwähnt:

Fünf Patienten beobachteten ihr störendes Verhalten, das während der Selbstbeobachtung zurückging. Eine Unterbrechung des Selbstberichts ließ die Frequenz wieder steigen; erneute Selbstbeobachtung führte zu weiterem Abfall der Häufigkeit. Wie in anderen Untersuchungen dieser Art fehlt auch bei Maletzky eine Angabe zur Verlässlichkeit der Selbstbeobachtung. Diese lassen sich in einer Studie von Herbert & Baer (1972) aus den Angaben externer Beobachter in den Familien der untersuchten Mütter und Kinder gewinnen, die dadurch zu einem *within-subject experimental design* wurde, daß eine Umkehrphase (reversal) eingeführt wurde, in der die Registrierung des eigenen Verhaltens durch die Mütter unterbrochen wurde. Einen Vergleich zwischen verschiedenen Vpn, also ein *between-subjects experimental design* finden wir in der Arbeit von Johnson & White (1971). Sie untersuchten drei Gruppen von College-Studenten. Die erste Gruppe beobachtete selbst ihre Studienaktivitäten, die zweite ihre Zeit, die sie mit Freundinnen verbrachten, die dritte Gruppe wurde nicht zu Selbstberichten angehalten. Abhängige Variable war die wöchentlich erhobene Benotung der Studienleistung. Darin zeigten sich signifikante Unterschiede zwischen der ersten und dritten Gruppe zugunsten der ersten, während sich die zweite Gruppe nicht überzufällig von den beiden anderen unterschied. Das Erstellen von Selbstberichten über eine bestimmte Verhaltensweise (dating) kann demnach auch zu Effekten auf eine andere Verhaltensweise (Studiengewohnheiten) führen, also generalisieren, und dann die abhängige Variable indirekt beeinflussen.

Die Stärke der Reaktivität hängt von der Bewertung des beobachteten Verhaltens und der Motivation, es zu ändern ab: Wird das Verhalten positiv eingeschätzt, steigt seine Frequenz, bei negativer Bewertung sinkt sie (Brodén et al. 1971, Kazdin 1974, Cavior & Marabotto 1976). Kanfer (1970) erklärt die Wirkung der Reaktivität als einen Prozeß mit drei Stadien: Im ersten Schritt erfolgt die Selbstbeobachtung, im zweiten bewertet die Person selbst ihr Verhalten nach ihren eigenen Normen, danach führt im dritten Schritt eine positive Bewertung zum Anstieg, die negative zum Abfall der Auftretenshäufigkeit. Diese von Ciminero et al. als „mediational explanation“ - wegen der Vermittlung über die Bewertung - bezeichnete Erklärung wird von ihnen mit der „operant explanation“ von Rachlin (1974) kontrastiert. Nach Rachlin haben die Folgen, die man sich selbst verschafft, nicht die Funktion von Verstärkern, sondern von Hinweisreizen. Diese Hinweisreize wirken als Signale für das, was schließlich an externen Konsequenzen zu erwarten sei, wenn man sich auf ein bestimmtes Verhalten einläßt. Selbstberichte von eigenem Verhalten schlosse dann tendenziell zumindest auch Beobachten der Folgen dieses eigenen Verhaltens mit ein.

3.5.3 Einflüsse bestehender Erwartungen der Beobachter

Einige der in den vorausgegangenen Abschnitten geschilderten Ergebnisse lassen sich integrieren, wenn man annimmt, daß Beobachter *Erwartungen entwickeln*, denen sie sich in ihrem Beurteilungsverhalten anpassen. Kent & Foster (1977) bezeichnen die Kovariation von geäußelter Erwartung über vermutlich auftretendes Verhalten und vom gleichen Beobachter berichtetem Verhalten als *expectation bias* (auch als Rosenthal-Effekt bekannt). Die Erwartungen der Beobachter beziehen sich unter anderem auf Art und Ausprägung des vermutlich auftretenden Verhaltens der zu beobachtenden Personen, auf Beurteilungsverhalten etwaiger Mitbeobachter und auf Kontrollverhalten der Forscher. Tritt höhere Verhaltensvariation des Beobachtungsobjektes auf als die Beobachter erwarten, oder zeigt sich das zu Beobachtende in einer für den Beobachter neuen Umgebung, so könnten Kategoriensystem und Beobachtungsinstruktion neu interpretiert werden müssen (was die hierbei gefundenen, vorübergehenden Reliabilitätsverminderungen erklären kann). Lernt der Beobachter, was er von Mitbeobachtern zu erwarten hat, so kann er sich anpassen, was zu der meist erwünschten Steigerung der Übereinstimmung führt. Auf Motivation, Aufmerksamkeit und Konzentration schließlich wirkt eine Kontrollerwartung zugunsten verbesserter Verlässlichkeit, während die oft erhöhten Anforderungen an Gedächtnis, Aufmerksamkeit und ähnlichen kognitiven Leistungen, welche die Einsatzsituation oft kennzeichnen, auch die Verlässlichkeit der Beobachtungen beeinträchtigen können.

Schon 1961, vor Rosenthal also, zeigten Azrin et al., daß Beurteilungen, die Beobachter über Meinungsäußerungen abgaben, mit ihren Erwartungen systematisch variierten, welche Meinung von den Vpn geäußert werden würde. Auf welche Schwierigkeiten Studien zum Rosenthal-Effekt stoßen, wird aus der Dissertation von Kent (1972, hier nach Kent & Foster 1977) deutlich: Zunächst wurden alle 40 Vpn gemeinsam geschult, und zwar 40 Stunden lang, verteilt über eine Periode von fünf Wochen anhand von Aufzeichnungen mit Videotape. Wie sich zeigt, ist gemeinsame Schulung wichtig, um nicht zu Unterschieden zwischen den Beobachtern zu kommen, die schon vor ihrer Zuteilung zu Versuchsbedingungen bestehen und Vergleiche zwischen den Versuchsbedingungen konfundieren. Gegen Ende der Schulung stellte sich heraus, daß sich die Übereinstimmung von $r = 0,60$ zwischen den Beobachtern nicht mehr verbessern ließ - zur Berechnung wurde jeder Beobachter mit einem aus den übrigen 39 zufällig herausgegriffenen verglichen. Danach wurden die Beobachter in zufälligen Gruppen zu je 5 Personen den 8 Versuchsbedingungen mit unterschiedlichen Ergebniserwartungen zugeteilt. Schon nach drei Tagen stieg die mittlere Verlässlichkeit innerhalb dieser Fünfergruppe auf 0,7. Zugleich entwickelten sich, und zwar vor jeder experimentellen Variation Unterschiede zwischen den Gruppen. Dieser Effekt wurde als „consensual observer drift“ (s. Johnson & Bolstad 1975) bezeichnet und führt Kent & Foster (1977, S. 283) zu der Warnung:

„... any study in which groups of observers collect data only within a particular class-room or treatment condition may badly confound differences in use of the behavioral code with the variables under investigation.“

Um Schwierigkeiten mit Vortreatment-Unterschieden auszuräumen, gingen Kent et al. (1974) so vor: Zehn Paare von Beobachtern (undergraduates) wurden so lange trainiert, bis jeder mit seinem Partner hinreichend hohe Übereinstimmung zeigte. Dann wurde at random je ein Mitglied eines Paares einer der beiden Versuchsbedingungen zugeteilt und beobachtete getrennt Videotapes, die angeblich teils das störende Verhalten eines Kindes vor, teils während der Behandlung zeigten. Der Versuchsleiter in der einen Experimentalbedingung informierte die Beobachter dahingehend, das Störverhalten werde während der Behandlung abnehmen, während in der anderen Bedingung - wahrheitsgemäß - informiert wurde, es werde sich keine Änderung zeigen. In der Auswertung ergab eine globale Frage, ob Veränderung eingetreten sei, den Rosenthal-Effekt; wertete man jedoch die detaillierten Verhaltensprotokolle zu neun operational definierten Kategorien aus, so zeigten sich keine erwartungsbedingten Unterschiede. Kent & Foster halten es für möglich, daß Erwartungseffekte insbesondere bei globalen Bewertungen, nicht jedoch bei spezifischen Verhaltensbeschreibungen auftreten. Allerdings zeigten sich in einer Studie von O'Leary et al. (1975) auch bei relativ explizit definierten Kategorien Effekte von Äußerungen von Versuchsleitern, die eine Erwartung bewirken sollten.

Literatur

- Ach, N. 1905. über die Willenstätigkeit und das Denken. Göttingen: Vandenhoeck & Ruprecht.
- Adair, J. & Spinner, B. 1979. Subjects' access to cognitive processes: Demand characteristics and verbal report. Unpublished manuscript, University Manitoba.
- Azrin, N. H., Holz, W., Ulrich, R. & Goldiamond, I. 1961. The control of the content of conversation through reinforcement. *Journal of experimental analysis of behavior*, **4**, 25-30.
- Bacon, F. 1974. Neues Organ der Wissenschaften (1620). (Dtsch. von A. T.Brück). Darmstadt: Wissenschaftliche Buchgesellschaft.
- Bakan, D. 1959. A reconsideration of the problem of introspection. *Psychological Bulletin*, **51**, 105-118.
- Barker, R. 1968. Ecological psychology: concepts and methods for studying the environment of human behavior. Stanford: Stanford Univ.Press.
- Barker, R. & Schoggen, P. 1973. Qualities of community life. San Francisco: Jossey-Bass.

- Bern, D. J. 1965. An experimental analysis of self-persuasion. *Journal of Experimental Social Psychology*, 1, 199-218.
- Bern, D. J. 1966. Inducing belief in false confessions. *Journal of Personality and Social Psychology*, 3, 707-710.
- Bern, D. J. 1972. Self-perception theory. In: Berkowitz, L. (ed.): *Advances in experimental social psychology*. Vol.6, New York: Academic Press, 1-62.
- Berman, J. S. & Kenny, D. A. 1976. Correlational bias in observer ratings. *Journal of Personality and Social Psychology*, 34, 263-273.
- Berman, J. S. & Kenny, D. A. 1977. Correlational bias: Not gone and not to be forgotten. *Journal of Personality and Social Psychology*, 35, 882-887.
- Birdwhistell, R. L. 1970. *Kinesics and context*. Philadelphia: Univ. of Philadelphia Press.
- Block, J. 1977. Correlational bias in observer ratings: Another perspective on the Berman and Kenny study. *Journal of Personality and Social Psychology*, 35, 873-880.
- Blumenthal, A. L. 1975. A reappraisal of Wilhelm Wundt: *American Psychologist*, 30, 1081-1088.
- Borgatta, E. F., Cottrell, L. S. & Mann, J. H. 1958. The spectrum of individual interaction characteristics: An inter-dimensional analysis. *Psychological Reports*, 4, 279-319.
- Braun, P. 1978. Verhaltenstherapeutische Diagnostik. In: L. Pongratz (ed.): *Klinische Psychologie (Handbuch der Psychologie. Bd. 8/2)*. Göttingen: Hogrefe, 1649-1725.
- Broden, M., Hall, R. V. & Mitts, B. 1971. The effect of self-recording on the classroom behavior of two eighth-grade students. *Journal of Applied Behavior Analysis*, 4, 191-199.
- Brown, A. L. 1978. Knowing when, where, and how to remember: A problem of metacognition. In: R. Clark (ed.): *Advances in instructional psychology*, Vol. I. Hillsdale, N.J.: Erlbaum.
- Bruner, J. S. & Tagiuri, R. 1954. The perception of people. In: Lindzey, G. (ed.): *Handbook of social psychology*. Vol.2, Cambridge, Mass.: Addison-Wesley, 634-654.
- Brunswik, E. 1952. The conceptual framework of psychology. *International Encyclopedia of Unified Science*, 1, No.10.
- Brunswik, E. 1955. Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62, 193-217.
- Brunswik, E. 1956. *Perception and the representative design of experiments*. Berkeley: Univ. of California Press.
- Castellan, N. J., Jr. 1973. Comments on the „lens model“ equation and the analysis of multiple-tue judgment tasks. *Psychometrika*, 38, 87-100.
- Cattell, R. B. 1957. *Personality and motivation: Structure and measurement*. Yonkers, N. Y.: World Book.

- Cavanaugh, J. C. & Perlmutter, M. 1980. Metamemory - a critical examination. Unpublished paper (draft). Institute of Child Development. University of Minnesota.
- Cavior, N. & Marabotto, C. 1976. Monitoring verbal behaviors in dyadic interaction. *Journal of Consulting and Clinical Psychology*, 44, 68-76.
- Chapman, L. J. & Chapman, J. P. 1967. The genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology*, 72, 193-204.
- Chapman, L.J. & Chapman, J. P. 1969. Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, 74, 271-280.
- Chi, P. L. 1937. Statistical analysis of personality. *Journal of Experimental Education*, 5, 229-245.
- Ciminero, A. R. 1977. Behavioral assessment: An overview. In: Ciminero, A. R., Calhoun, K. S. & Adams, H. E. (eds): *Handbook of behavioral assessment*. New York: Wiley, 3-13.
- Ciminero, A. R., Nelson, R. O. & Lipinski, D. P. 1977. Self-monitoring procedures. In: Ciminero, A. R., Calhoun, K. S. & Adams, H. E. (eds): *Handbook of behavioral assessment*. New York: Wiley, 195-232.
- Claparède, E. 1965. Die Entdeckung der Hypothese. In: C. F. Graumann (ed.): *Denken*. Köln: Kiepenheuer & Witsch, 109-115.
- Cohen, R. 1971. An investigation of the diagnostic processing of contradictory information. *European Journal of Social Psychology*, 1, 475-492.
- Cohen, R. & Schümer, R. 1968. Eine Untersuchung zur sozialen Urteilsbildung. I. Die Verarbeitung von Informationen unterschiedlicher Konsonanz. *Archiv für die gesamte Psychologie*, 120, 151-179.
- Comte, A. 1949. *Cours de philosophie positive (1830-1842)*. 2 Bände. Paris: Garnier.
- Coombs, C. H., Dawes, R. M. & Tversky, A. 1970. *Mathematical Psychology*. Englewood Cliffs, N. J.: Prentice-Hall.
- Cronbach, L. J. 1955. Processes affecting scores of „understanding of others“ and „assumed similarity“. *Psychological Bulletin*, 52, 177—193.
- D'Andrade, R. G. 1965. Trait psychology and componential analysis. *American Anthropologist*, 67, 215-228.
- D'Andrade, R. G. 1974. Memory and the assessment of behavior. In: Blalock, H. M. Jr. (ed.): *Measurement in the social sciences*. Chicago: Aldine, 159-186.
- Danziger, K. 1980. The history of introspection reconsidered. *Journal of the History of the Behavioral Sciences*, 16, 241-262.
- Dörner, D. 1974. *Die kognitive Organisation beim Problemlösen*. Bern: Huber.
- Duncan, S. 1969. Nonverbal communication. *Psychological Bulletin*, 72, 118-137.
- Duncker, K. 1926. A qualitative (experimental and theoretical) study of productive thinking (solving of comprehensible problems). *The Pedagogical Seminary and Journal of Genetic Psychology*, 33, 642-708.
- Duncker, K. 1966. *Zur Psychologie des produktiven Denkens (1935)*. Berlin: Springer.

- Egan, J. P. 1975. Signal detection theory and ROC analysis. New York: Academic Press.
- Ericsson, K. A. & Simon, H. A. 1980. Verbal reports as data. *Psychological Review*, 37, 215-251.
- Feger, H. 1972. Skalierte Informationsmenge und Eindrucksurteil. Bern: Huber.
- Feger, H. & Feger, B. 1969 (a). Beiträge zur inhaltsanalytischen Untersuchung von Entscheidungen. Teil I: Methode und Vergleich der Materialstichproben. *Archiv für die gesamte Psychologie*, 121, 205-232.
- Feger, H. & Feger, B. 1969 (b). Beiträge zur inhaltsanalytischen Untersuchung von Entscheidungen. Teil II: Kontingenzanalyse und Paralleluntersuchung. *Archiv für die gesamte Psychologie*, 121, 233-254.
- Finke, R. A. & Kosslyn, S. M. 1980. Mental imagery acuity in the peripheral visual field. *Journal of Experimental Psychology: Human Perception and Performance*, 6, 126-139.
- Fixen, D. L., Phillips, E. L. & Wolf, M.M. 1972. Achievement place: The reliability of self-reporting and peer-reporting and their effects on behavior. *Journal of Applied Behavior Analysis*, 5, 19-30.
- Flavell, J. H. 1976. Metacognitive aspects of problem-solving. In: L. B. Resnick (ed.): *The nature of intelligence*. Hillsdale, N. J.: Erlbaum.
- Flavell, J. H. 1979. Monitoring social-cognitive enterprises: Something else that may develop in the area of social cognition. Paper presented for the Social Science Research Council Committee on Social and Affective Development During Childhood, January.
- Gibson, J. J. 1950. *The perception of the visual world*. Boston: Houghton Mifflin, (Deutsch: *Die Wahrnehmung der visuellen Welt*. Weinheim: Beltz, 1973).
- Giorgi, A. 1967. A phenomenological approach to the problem of meaning and social learning. *Review of Existential Psychology and Psychiatry*, 7, 106-118.
- Goldfried, M. R. & Kent, R. N. 1972. Traditional versus behavioral assessment: A comparison of methodological and theoretical assumptions. *Psychological Bulletin*, 77, 409-420.
- Goldfried, M. R. & Linehan, M. M. 1977. Basic issues in behavioral assessment. In: Ciminero, A. R., Calhoun, K. S. & Adams, H. E. (eds): *Handbook of behavioral assessment*. New York: Wiley, 15-46.
- Graumann, C. F. 1966. Bewußtsein und Bewußtheit - Probleme und Befunde der psychologischen Bewußtseinsforschung. In: W. Metzger (ed.): *Wahrnehmung und Bewußtsein* (Handbuch der Psychologie, I, 1). Göttingen: Hogrefe, 79-127.
- Graumann, C. F. 1978. Wahrnehmung und Beurteilung der anderen und der eigenen Person. In: A. Heigl-Evers (ed.): *Lewin und die Folgen* (Die Psychologie des 20. Jahrhunderts, Bd. VIII). Zürich: Kindler, 154-183.
- Graumann, C. F. 1980. Experiment, Statistik, Geschichte - Wundts erstes Heidelberger Programm einer Psychologie. *Psychologische Rundschau*, 31, 73-83

- (Auch in: W. G. Bringmann & R. D. Tweney (eds): *Wundt Studies*. Toronto: Hogrefe, 1980, 33-41).
- Greeno, J. G. 1968. *Elementary theoretical psychology*. Reading, Mass.: Addison-Wesley.
- Grüner, K.-W. 1974. *Beobachtung*. Stuttgart: Teubner (Studienskripten).
- Guilford, J. P. 1954. *Psychometric methods*. New York: McGraw-Hill, (2. Aufl.).
- Gulliksen, H. 1950. *Theory of mental tests*. New York: Wiley.
- Gulliksen, H. 1968. Methods for determining equivalence of measures. *Psychological Bulletin*, 70, 534-544.
- Gurwitsch, A. 1966. *Studies in phenomenology and psychology*. Evanston, Ill.: Northwestern.
- Hagen, R. L., Craighead, W. E. & Paul, G. L. 1975. Staff reactivity to evaluative behavioral observations. *Behavior Therapy*, 6, 201-205.
- Hart, J. T. 1965. Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, 56, 208-216.
- Hart, J. T. 1966. Methodological note on feeling-of-knowing experiments. *Journal of Educational Psychology*, 57, 347-349.
- Hart, J. T. 1967. Second-try recall, recognition, and the memory-monitoring process. *Journal of Educational Psychology*, 58, 193-197.
- Herbert, E. W. & Baer, D. M. 1972. Training parents as behavior modifiers: Self-recording of contingent attention. *Journal of Applied Behavioral Analysis*, 5, 139-149.
- Holzkamp, K. 1980. Zu Wundts Kritik an der experimentellen Erforschung des Denkens. In: Wilhelm Wundt - Progressives Erbe, Wissenschaftsentwicklung und Gegenwart. Leipzig: Karl-Marx-Universität, 141-153.
- Ingleby, J. D. 1974. Further studies of the human observer as a statistical decision maker. *Organizational Behavior and Human Performance*, 12, 299-314.
- James, W. 1950. *The principles of psychology*. 2 vols. (1890). New York: Dover.
- Johnson, D. M. 1963. Reanalysis of experimental halo effects. *Journal of Applied Psychology*, 47, 46-47.
- Johnson, S. M. & Bolstad, O. D. 1975. Reactivity to home observation: A comparison of audio recorded behavior with observers present or absent. *Journal of Applied Behavioral Analysis*, 8, 181-185.
- Johnson, S. M. & Lobitz, G. K. 1974. Parental manipulations of child behavior in home observations. *Journal of Applied Behavior Analysis*, 7, 23-32.
- Johnson, S. M. & White, G. 1971. Self-observation as an agent of behavioral change. *Behavior Therapy*, 2, 488-497.
- Johnson, S. M., Christensen, A. & Bellamy, G. T. 1976. Evaluation of family intervention through unobtrusive audio recordings: Experiences in bugging children. *Journal of Applied Behavioral Analysis*, 9, 213-219.

- Jones, R. R., Reid, J. B. & Patterson, G. R. 1975. Naturalistic observation in clinical assessment. In: McReynolds, P. (ed.): *Advances in psychological assessment*. Vol. 3. San Francisco: Jossey-Bass, 42-95.
- Kanfer, F. H. 1970. Self-monitoring: Methodological limitations and clinical applications. *Journal of Consulting and Clinical Psychology*, 35, 148-152.
- Kanfer, F. H. 1975. Self-management methods. In: F. H. Kanfer & A. P. Goldstein, *Helping People Change*. New York: Pergamon.
- Kant, I. 1800. *Anthropologie in pragmatischer Hinsicht*. 2. Aufl. Königsberg: Nicolovius.
- Kant, I. 1903. *Metaphysische Anfangsgründe der Naturwissenschaft (1786)*. In: I. Kant: *Gesammelte Schriften*, Bd. 4. Berlin: Reimer.
- Katz, D. 1911. Die Erscheinungsweisen der Farben und ihre Beeinflussung durch die individuelle Erfahrung. *Zeitschrift für Psychologie, Erg.band 7*.
- Katz, D. 1929. Der Aufbau der Tastwelt. *Zeitschrift für Psychologie, Erg.band 11*.
- Kazdin, A. E. 1974. Reactive self-monitoring: The effects of response desirability, goal setting, and feedback. *Journal of Consulting and Clinical Psychology*, 42, 704-716.
- Kent, R. N. 1972. Expectancy bias in behavioral observation. Unpubl. doc.diss., State Univ. of New York, Stony Brook, New York.
- Kent, R. N. & Foster, S. L. 1977. Direct observational procedures: Methodological issues in naturalistic settings. In: Ciminero, A. R., Calhoun, K. S. & Adams, H. E. (eds): *Handbook of behavioral assessment*. New York: Wiley, 279-328.
- Kent, R. N., O'Leary, K. D., Diamant, C. & Dietz, A. 1974. Expectation biases in observational evaluation of therapy change. *Journal of Consulting and Clinical Psychology*, 42, 774-780.
- Köhler, W. 1921. *Intelligenzprüfungen an Menschenaffen*. Berlin: Springer.
- Lachman, J. L. & Lachman, R. 1980. Age and the actualization of world knowledge. In: L. W. Poon et al. (eds): *New directions in memory and aging*. Hillsdale, N.J.: Erlbaum, 285-313.
- Lambert, W. W. 1960. Interpersonal behavior. In: Mussen, P. H. (ed.): *Methods in child development*. New York: Wiley.
- Lay, C. H. & Jackson, D. N. 1969. Analysis of the generality of traitinferential relationship. *Journal of Personality and Social Psychology*, 12, 12-21.
- Lieberman, D. A. 1979. A (limited) call for a return to introspection. *American Psychologist*, 34, 319-333.
- Linschoten, J. 1959. *Op weg naar een fenomenologische psychologie*. Utrecht: Bijleveld, (übers. von F. Mönks: *Auf dem Wege zu einer phänomenologischen Psychologie*. Berlin: de Gruyter, 1961).
- Lipinski, D. P. & Nelson, R. O. 1974. The reactivity and unreliability of self-recording. *Journal of Consulting and Clinical Psychology*, 42, 118-123.
- Lipinski, D. P., Black, J. L., Nelson, R. O. & Ciminero, A. R. 1975. The influence of

- motivational variables on the reactivity and reliability of self-recording. *Journal of Consulting and Clinical Psychology*, 43, 637-646.
- Longabaugh, R. 1980. The systematic observation of behavior in naturalistic settings. In: Triandis, H. C. & Berry, J. W. (eds): *Handbook of cross-cultural psychology*. Vol. 2: Methodology. Boston: Allyn & Bacon, 57-126.
- Lorr, M. & McNair, D. M. 1965. Expansion of the interpersonal behavior circle. *Journal of Personality and Social Psychology*, 2, 823-830.
- Lüer, G. 1973. *Gesetzmäßige Denkabläufe beim Problemlösen*. Weinheim: Beltz.
- Maletzky, B. M. 1974. Behavior recording as treatment: A brief note. *Behavior Therapy*, 5, 107-111.
- Mann, R. D. 1959. The relation between personality characteristics and individual performance in small groups. Ph. D. dissertation, Univ. of Michigan.
- Mash, E. J. & McElwee, J. D. 1974. Situational effects on observer accuracy: Behavior predictability, prior experience, and complexity of coding categorie. *Child Development*, 45, 367-377.
- Mash, L. J. & Makohoniuk, G. 1975. The effects of prior information and behavioral predictability on observer accuracy. *Child Development*, 46, 513-519.
- McNicol, D. 1972. *A primer of signal detection theory*. London: Allen & Unwin.
- Meichenbaum, D., Burland, S., Gruson, L. & Cameron, R. 1979. Metacognitive assessment. Paper presented at the Conference on the Growth of Insight. Wisconsin Research and Development Center, October.
- Meichenbaum, D. & Butler, L. 1980. Cognitive ethology: Assessing the streams of cognition and emotion. In: K. Blankstein, P. Pliner & J. Polivy (eds): *Advances in the study of communication and affect: Assessment and modification of emotional behavior*. Vol. 6. New York: Plenum.
- Metge, Anneros. 1980. Zum Problem der Selbstbeobachtung bei Wundt. In: *Wilhelm Wundt - Progressives Erbe. Wissenschaftsentwicklung und Gegenwart*. Leipzig: Karl-Marx-Universität, 183-190.
- Metzger, W. 1954. *Psychologie*. 2. Aufl. Darmstadt: Steinkopf.
- Mitchell, D. B. & Richman, C. L. 1980. Confirmed reservations: Mental travel. *Journal of Experimental Psychology: Human Perception and Performance*, 6, 58-66.
- Moos, R. H. 1968. Behavioral effects of being observed: Reactions to a wireless radio transmitter. *Journal of Consulting and Clinical Psychology*, 32, 383-388.
- Mulaik, A. 1964. Are personality factors rater's conceptual factors? *Journal of Consulting Psychology*, 28, 506-511.
- Natsoulas, T. 1970. Concerning introspective 'knowledge'. *Psychological Bulletin*, 73, 89-111.
- Natsoulas, T. 1978. Residual subjectivity. *American Psychologist*, 33, 269-283.
- Nelson, R. O., Lipinski, D. P. & Black, J. L. 1975. The effects of expectancy on the reactivity of self-recording. *Behavior Therapy*, 6, 337-349.

- Newcomb, T. M. 1931. An experimental design to test the validity of a rating technique. *Journal of Educational Psychology*, 22, 279-289.
- Newton, D. 1973. Attribution and the unit of perception of ongoing behavior. *Journal of Personality and Social Psychology*, 28, 28-38.
- Newton, D. 1976. Foundations of attribution: The unit of perception of ongoing behavior. In: Harvey, J., Ickes, W. & Kidd, R. (eds): *New directions in attribution research*. Hillsdale, N. J.: Erlbaum, 223-247.
- Newton, D. & Engquist, G. 1976. The perceptual organization of ongoing behavior. *Journal of Experimental Social Psychology*, 12, 436-450.
- Newton, D. & Rindner, R. J. 1979. Variation in behavior perception and ability attribution. *Journal of Personality and Social Psychology*, 37, 1847-1858.
- Newton, D., Engquist, G. & Bois, J. 1977. The objective basis of behavior units. *Journal of Personality and Social Psychology*, 35, 847-862.
- Nisbett, R. E. & Wilson, T. D. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84, 231-259.
- Nuttin, J. 1955. Consciousness, behavior, and personality. *Psychological Review*, 62, 349-355.
- O'Leary, K. D., Kent, R. N. & Kanowitz, J. 1975. Shaping data collection congruent with experimental hypotheses. *Journal of Applied Behavior Analysis*, 8, 43-51.
- Passini, F. T. & Norman, W. T. 1966. A universal conception of personality structure? *Journal of Personality and Social Psychology*, 4, 44-49.
- Pastore, R. E. & Scheirer, C. J. 1974. Signal detection theory: Considerations for general application. *Psychological Bulletin*, 81, 945-958.
- Peterson, D. R. 1968. *The clinical study of social behavior*. New York: Appleton-Century-Crofts.
- Peterson, W. W., Birdsall, T. G. & Fox, W. C. 1954. The theory of Signal detectability. *Institute of Radio Engineers Transactions*, PGIT-4, 171-212.
- Pilkington, C. W. & Glasgow, W. D. 1967. Towards a rehabilitation of introspection as a method in psychology. *Journal of Existentialism*, 7, 329-350.
- Podgorny, P. & Shepard, R. N. 1978. Functional representations common to visual perception and imagination. *Journal of Experimental Psychology: Human Perception and Performance*, 4, 21-35.
- Postman, L. & Tolman, E. C. 1959. Brunswik's probabilistic functionalism. In: Koch, S. (ed.): *Psychology: A study of a science*. Vol. I. New York: McGraw-Hill, 502-564.
- Price, R. H. 1966. Signal detection methods in personality and perception. *Psychological Bulletin*, 66, 55-62.
- Rachlin, H. 1974. Self-control. *Behaviorism*, 2, 94-107.
- Radford, J. 1974. Reflections on introspection. *American Psychologist*, 29, 245-250.
- Redfield, J. & Paul, G. L. 1976. Bias in behavioral observation as a function of observer

- familiarity with subjects and typicality of behavior. *Journal of Consulting and Clinical Psychology*, 44, 156.
- Reid, J. B. 1970. Reliability assessment of observational data: A possible methodological problem. *Child Development*, 41, 1143-1150.
- Richardson, J. T. E. 1980. *Mental imagery and human memory*. London: Macmillan.
- Roberts, R. R. Jr. & Renzaglia, G. A. 1905. The influence of tape recording on counseling. *Journal of Counseling Psychology*, 12, 10-16.
- Rohracher, H. 1963. *Einführung in die Psychologie*. 8. Aufl. Wien: Urban & Schwarzenberg.
- Romanczyk, R. G., Kent, R. N., Diament, C. & O'Leary, K. D. 1973. Measuring the reliability of observational data: A reactive process. *Journal of Applied Behavior Analysis*, 6, 175-184.
- Rosenthal, R. 1976. *Experimenter effects in behavioral research*. New York: Appleton-Century-Crofts, 1966; enlarged edition: New York: Irvington Publ.
- Rosenthal, R. & Rosnow, R. L. (eds) 1969. *Artifact in behavioral research*. New York: Academic Press.
- Rudinger, G. & Feger, H. 1970. Die Beurteilung formaler Persönlichkeitsmerkmale durch Rating-Skalen: Eine Generalisierbarkeitsstudie. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 2, 96-112.
- Rugg, H. 1922. Is the rating of human Character practicable? *Journal of Educational Psychology*, 13, 30-42.
- Schapp, W. 1976. *Beiträge zur Phänomenologie der Wahrnehmung*. Wiesbaden: Heymann.
- Schefflen, A. E. 1975. *How behavior means*. New York: Aronson.
- Schümer, R. 1971. Eine experimentelle Untersuchung zur sozialen Eindrucksbildung. *Zeitschrift für Sozialpsychologie*, 2, 92-108.
- Shepard, R. N. 1978. The mental image. *American Psychologist*, 33, 125-137.
- Shepard, R. N. & Chipman, S. 1970. Second order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, 1, 1-17.
- Shepard, R. N. & Metzler, J. 1971. Mental rotation of three-dimensional objects. *Science*, 171, 701-703.
- Simkins, L. 1971. The reliability of self-recorded behaviors. *Behavior Therapy*, 2, 83-87.
- Skinner, B. F. 1945. The operational analysis of psychological terms. *Psychological Review*, 52, 270-277.
- Skinner, B. F. 1953. *Science and human behavior*. New York: Macmillan, (Deutsch: *Wissenschaft und menschliches Verhalten*. München: Kindler, 1973).
- Skinner, B. F. 1957. *Verbal behavior*. New York: Appleton.
- Skinner, B. F. 1963. Behaviorism at fifty. *Science*, 140, 951-958.
- Skinner, B. F. 1974. *About behaviorism*. New York: Knopf.

- Stanley, J. C. 1961. Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. *Psychometrika*, 26, 205-219.
- Starr, D. J. & Katkin, E. S. 1969. The clinician as aberrant actuary: Illusory correlation and the Incomplete Sentences Blank. *Journal of Abnormal Psychology*, 74, 670-675.
- Stern, W. 1911. *Die Differentielle Psychologie in ihren methodischen Grundlagen*. Leipzig: Barth.
- Steward, T. R. 1976. Components of correlation and extensions of the lens model equation. *Psychometrika*, 41, 101-120.
- Stricker, L. J., Jacobs, P. I. & Kogan, N. 1974. Trait interrelations in implicit personality theory and questionnaire data. *Journal of Personality and Social Psychology*, 30, 198-207.
- Swets, J. A. (ed.) 1964. *Signal detection and recognition by human observers*. New York: Wiley.
- Taplin, P. S. & Reid, J. B. 1973. Effects of instructional set and experimenter influence on observer reliability. *Child Development*, 44, 547-554.
- Thorndike, E. L. 1920. A constant error in psychological ratings. *Journal of Applied Psychology*, 4, 25-29.
- Thomae, H. 1960. *Der Mensch in der Entscheidung*. München: Barth.
- Titchener, E. B. 1914. *A primer in psychology*. New York: Macmillan.
- van der Kamp, L. J. T. & Mellenbergh, G. J. 1976. Agreement between raters. *Educational and Psychological Measurement*, 36, 311-317.
- van Meter, D. & Middleton, D. 1954. Modern statistical approaches to reception in communication theory. *Institute of Radio Engineers Transactions*, PGIT-4, 119-145.
- Webb, E. J., Campbell, D. T., Schwartz, R. D. & Sechrest, L. 1966. Unobtrusive measures. *Nonreactive research in the social sciences*. Chicago: Rand McNally, (7. Aufl. 1971; deutsch als: *Nichtreaktive Meßverfahren*. Weinheim: Beltz, 1975).
- Wells, F. L. 1907. A statistical study of literary merit. *Archives of Psychology*, 1 (7).
- Wertheimer, M. 1912. Experimentelle Studien über das Sehen von Bewegungen. *Zeitschrift für Psychologie*, 61, 161-265.
- Werts, C. E., Jöreskog, K. G. & Linn, R. L. 1976. Analyzing ratings with correlated intrajudge measurement errors. *Educational and Psychological Measurement*, 36, 319-328.
- White, P. 1980. Limitations on verbal reports of internal events: A refutation of Nisbett and Wilson and of Bern. *Psychological Review*, 87, 105-112.
- Whiting, B. B. & Whiting, J. W. 1975. *Children in six cultures: a psycho-cultural analysis*. Cambridge, Mass.: Harvard Univ. Press.
- Wilder, D. 1978. Effects of predictability on units of perception and attribution. *Personality and Social Psychology Bulletin*, 4, 281-284.

- Wundt, W. 1862. Beiträge zur Theorie der Sinneswahrnehmung. Leipzig/Heidelberg: Winter.
- Wundt, W. 1863. Vorlesungen über die Menschen- und Thierseele. 2 Bde. Leipzig: Voß.
- Wundt, W. 1885. Die Aufgaben der experimentellen Psychologie (1882). In: W. Wundt: Essays. Leipzig: Engelmann.
- Wundt, W. 1888. Selbstbeobachtung und innere Wahrnehmung. Philosophische Studien, 4, 292-309.
- Wundt, W. 1907. über Ausfrageexperimente und über Methoden zur Psychologie des Denkens. Psychologische Studien, 3, 301-360.

3. Kapitel

Das Q-Sort-Verfahren

Wolf-Rüdiger Minsel und Manfred Heinz

1. Zur Einordnung des Q-Sort-Verfahrens

Das Q-Sort-Verfahren, kurz Q-Sort, stellt als Datenerhebungsverfahren eine spezielle Forschungsmethode dar. Es bildet eine Datenbasis für Q-Korrelationen und für die faktorenanalytische Q-Technik (vgl. Mowrer, 1953).

Nach Cattell (1957) sind Q-Daten diejenige Gruppe von Daten zur Persönlichkeitsbeschreibung, die aus der Selbstbeurteilung des Individuums mit Hilfe von Fragebögen oder Interviews erhoben werden. Bei Q-Korrelationen werden n Individuen über m Merkmale korreliert, wobei $m > n$ sein sollte. Diese Korrelationsmethode geht auf Stephenson (vgl. 1953) zurück und unterscheidet sich in statistischer Hinsicht nicht von der R-Technik (Dorsch, 1976). Es werden dabei Personen korreliert und faktorisiert. Die extrahierten Faktoren sind als Typen interpretierbar.

Besondere Charakteristika des Q-Sort-Verfahrens sind:

- es ist ein Rating-Verfahren (Langer & Schulz v. Thun, 1974), speziell zur Persönlichkeitsbeschreibung
- die Items werden theoriegeleitet und für den Individualfall entwickelt
- das Individuum liefert ipsative Daten; d.h. es werden Aussagen darüber gemacht, welche Persönlichkeitsmerkmale individuell als stark oder schwach in Relation zu anderen Persönlichkeitsmerkmalen und nicht im Vergleich zu anderen Personen oder zu einer ‚Außennorm‘ wahrgenommen werden.
- für die Organisation der Antworten wird eine Häufigkeitsbesetzung der einzelnen Rating-Kategorien zumeist in Form einer Normalverteilung erzwungen.

Die genannten Charakteristika sind in der Literatur vielfältig diskutiert worden und führten zu zahllosen von der ursprünglichen Form abweichenden Neuentwicklungen. Ältere Übersichtsarbeiten liegen vor zur Q-Technik von Mowrer (1953) und zur Q-Methodologie von Wittenborn (1961).

2. *Beispiel eines Q-Sort-Verfahrens*

Der wohl bekannteste Q-Sort ist der California Q-Set (CQ-Set) von Block (1961). Zum Zwecke der exemplarischen Darstellung des Verfahrens entscheiden wir uns jedoch für den Butler & Haigh-Q-Sort (1954), da dieser das originäre Konzept repräsentiert, leicht modifiziert übersetzt und teststatistisch überprüft vorliegt (vgl. Schön, 1966; Frohburg, 1972).

Der Butler & Haigh-Q-Sort dient der Erfassung des Selbstkonzeptes in Form eines Selbst- und Idealbildes und dient u.a. zur Überprüfung der Auswirkungen klientenzentrierter Psychotherapie (vgl. Minsel & Bente, 1979). Dabei wird erwartet, daß im Falle erfolgreicher Psychotherapiegespräche eine Annäherung des Selbstbildes an das Idealbild eines Klienten erfolgt.

Butler & Haigh gehen dabei von folgenden Grundannahmen aus:

- Das Selbstbild besteht aus einem organisierten Satz von Konzepten, die ein Individuum sich selbst zuschreibt, wie etwa „ich bin...“, „ich habe . . .“ usw.
- Die Konzepte können als Aussagen formuliert und von dem Individuum auf ihre Gültigkeit hin beurteilt werden.
- Die unterschiedlichen Werte oder Bedeutungen, die den Konzepten zukommen, lassen sich auf einer Ordinalskala abbilden; dabei gibt das Individuum den Grad seiner Zustimmung zu jeder Aussage an.
- Neben dem Selbstbild hat jedes Individuum einen organisierten Satz von Konzepten darüber, wie es im Idealfall sein möchte. Auch das Idealbild ist in gleicher Form wie das Selbstbild darstellbar und einschätzbar.

Bei der Anwendung des Butler & Haigh Q-Sorts werden dem Beurteiler 74 Karten vorgelegt. Auf jeder Karte steht ein Item, wie

- „I. Ich setze oft auf's falsche Pferd
- 27. Ich habe vor sexuellen Kontakten Angst
- 35. Ich bin wertlos
- 38. Ich bin ein lebenswürdiger Mensch
- 45. Ich bin anders als andere Menschen
- 58. Ich bin ausgeglichen
- 74. Ich stelle an mich selbst strenge Anforderungen.“

Der Beurteiler wird instruiert, zwei Sortierungen vorzunehmen. In der ersten soll er die Karten mit folgender Instruktion so arrangieren, daß sie das Bild ergeben, das er bei sich als aktuell wahrnimmt (Selbstbild, SB):

„Sortieren Sie diese Karten bitte so, daß die angegebenen Eigenschaften Sie so beschreiben, wie Sie sich heute sehen - von denen, die Ihnen am wenigsten ähnlich sind bis zu denen, die Ihnen am meisten ähnlich sind.“

In der zweiten Sortierung soll der Beurteiler die Karten so arrangieren, daß sie das Bild wiedergeben, wie er am liebsten sein möchte (Idealbild, IB):

„Nun sortieren Sie die Karten bitte so, daß Sie Ihre Idealperson beschreiben - die Person, die Sie am liebsten sein möchten.“

Die einzelnen Aussagen (Items) können in neun Kategorien sortiert werden:

- Kategorie 1: Aussagen, die überhaupt nicht zutreffen
- Kategorie 2: Aussagen, die kaum zutreffen
- Kategorie 3: Aussagen, die wenig zutreffen
- Kategorie 4: Aussagen, die etwas zutreffen
- Kategorie 5: Aussagen, die mittelmäßig charakteristisch sind
- Kategorie 6: Aussagen, die schon eher zutreffen
- Kategorie 7: Aussagen, die stärker zutreffen
- Kategorie 8: Aussagen, die sehr zutreffen
- Kategorie 9: Aussagen, die besonders typisch sind und genau zutreffen

Um eine Quasi-Normalverteilung zu erhalten, ist festgelegt, wieviele Items in die einzelnen Kategorien sortiert werden dürfen.

Kategorie	1	2	3	4	5	6	7	8	9
Anzahl an Items pro Kategorie	3	6	9	12	14	12	9	6	3

Für den Beurteiler wird ein Korrelationskoeffizient zwischen den Differenzwerten, die sich aus den Kategorienwerten unter beiden Sortierungen ergeben, berechnet. Frohburg (1972, S. 86, 87) erleichtert die Systematisierung der Daten und die Berechnung des Korrelations-Koeffizienten durch die nachfolgenden Übersichten (Abb. 1, Tab. 1).

Tabelle 1: Errechnung des Korrelationskoeffizienten für den Q-Sort nach der Formel

$$r = 1 - \frac{\sum D^2}{592}$$

ΣD^2	r	ΣD^2	r	ΣD^2	r	ΣD^2	r	ΣD^2	r
1184	-1.00	947	-.60	710	-.20	473	.20	237	.60
1178	-.99	941	-.59	704	-.19	468	.21	231	.61
1172	-.98	935	-.58	699	-.18	462	.22	225	.62
1166	-.97	929	-.57	693	-.17	456	.23	219	.63
1160	-.96	924	-.56	687	-.16	450	.24	213	.64
1154	-.95	918	-.55	681	-.15	444	.25	207	.65
1148	-.94	912	-.54	675	-.14	438	.26	201	.66
1142	-.93	906	-.53	669	-.13	432	.27	195	.67
1136	-.92	900	-.52	663	-.12	426	.28	189	.68
1130	-.91	894	-.51	657	-.11	420	.29	184	.69
1124	-.90	888	-.50	651	-.10	414	.30	177	.70
1118	-.89	882	-.49	645	-.09	408	.31	172	.71
1112	-.88	876	-.48	639	-.08	403	.32	166	.72
1107	-.87	870	-.47	633	-.07	397	.33	160	.73
1101	-.86	864	-.46	628	-.06	391	.34	154	.74
1095	-.85	858	-.45	622	-.05	385	.35	148	.75
1089	-.84	852	-.44	616	-.04	379	.36	142	.76
1083	-.83	847	-.43	610	-.03	373	.37	136	.77
1077	-.82	841	-.42	604	-.02	367	.38	130	.78
1072	-.81	835	-.41	598	-.01	361	.39	124	.79
1066	-.80	828	-.40	592	-.00	355	.40	118	.80
1060	-.79	823	-.39	586	.01	349	.41	112	.81
1054	-.78	817	-.38	580	.02	343	.42	107	.82
1048	-.77	811	-.37	574	.03	337	.43	101	.83
1042	-.76	805	-.36	568	.04	332	.44	95	.84
1036	-.75	799	-.35	562	.05	326	.45	89	.85
1030	-.74	793	-.34	556	.06	320	.46	83	.86
1024	-.73	787	-.33	551	.07	314	.47	76	.87
1018	-.72	781	-.32	545	.08	308	.48	71	.88
1012	-.71	776	-.31	539	.09	302	.49	65	.89
1006	-.70	770	-.30	533	.10	296	.50	59	.90
1000	-.69	764	-.29	527	.11	290	.51	53	.91
995	-.68	758	-.28	521	.12	284	.52	47	.92
989	-.67	752	-.27	515	.13	278	.53	41	.93
983	-.66	746	-.26	509	.14	272	.54	35	.94
977	-.65	740	-.25	503	.15	266	.55	30	.95
971	-.64	734	-.24	497	.16	260	.56	24	.96
965	-.63	728	-.23	491	.17	254	.57	18	.97
959	-.62	722	-.22	485	.18	248	.58	12	.98
953	-.61	716	-.21	479	.19	242	.59	6	.99
									1.00

Umrechnung der Differenzwerte (D) in Korrelationskoeffizienten (r) (nur gültig für eine Q-Sortierung mit 74 Items, sortiert in 9 Kategorien in der Verteilung 3, 6, 9, 12, 14, 9, 6, 3) (nach: Frohburg, 1972, S. 86)

Neben der skizzierten Möglichkeit, eine Selbst- und Idealbild-Sortierung vorzunehmen, finden sich zahlreiche variierende Instruktionsformen. Als einschätzbar gilt alles, was den Bereich der personalen und interpersonalen Wahrnehmung betrifft. Allerdings wird der Q-Sort dann vielfach nicht mehr idio-graphisch, sondern nomothetisch verwendet. Ein Beispiel dafür liefert Hartley (1950). Die Autorin ließ die an ihrer Untersuchung beteiligten Personen neben der tradierten Selbst- vs Idealbild-Sortierung solche für das „unglückliche Selbst“ und die „normale Mutter“ durchführen. Im gleichen Sinne ließ auch Block (1961) die „normal angepaßte Person“ anhand seines CQ-Sets einschätzen.

3. Anwendung des Q-Sort-Verfahrens

Die Entwicklung des Verfahrens geht auf Stephenson (vgl. 1953) zurück. Für ihn bildete das Verfahren einen neuen Ausgangspunkt zur Persönlichkeitsbeschreibung. Mit dieser Rating-Methode erhob Stephenson Daten, die er mit Hilfe einer ‚inversen‘ faktorenanalytischen Technik verrechnete. Es wurden dabei Korrelationen zwischen Personen, bzw. zwischen den Daten einer Person, die unter verschiedenen Bedingungen gefunden wurden, anstelle von Korrelationen von Tests berechnet. Stephensons erklärte Zielsetzung war, den Individualfall zum Gegenstand der Faktorenanalyse zu machen.

Tatsächlich hat das Q-Sort-Verfahren jedoch ein sehr viel breiteres Anwendungsfeld gefunden. Das galt besonders in der 50er Jahren. In verschiedensten Formen’ unter unterschiedlichsten Bedingungen und Fragestellungen wurde das Verfahren angewandt. Einige Beispiele sollen das veranschaulichen. So wurde das Q-Sort-Verfahren benutzt

- zur Charakterisierung von Persönlichkeitstypen (vgl. Shontz, 1956, Nakhinsky, 1958; Block, 1961)
- zur Diagnostik psychischer Anpassung vs Fehlanpassung (vgl. Rogers & Dymond, 1954; Friedman, 1955; Chase, 1957; Cartwright 1957; Turner & Vanderlippe, 1958; Tobacyk, Broughton & Vaught, 1975)
- zur Überprüfung der Wirksamkeit psychotherapeutischer Interventionen (vgl. Fiedler, 1951; Rogers & Dymond, 1954; Shlien, 1964; Garfield & Prager, 1971; Sherry & Hurley, 1976)
- zur Untersuchung des Selbstkonzeptes unterschiedlicher Populationen, wie u.a. von Kindern, Jugendlichen, Stotterern, Schizophrenen (vgl. Fiedler, 1951; Caplan, 1957; Perkins, 1958; Rogers, 1958)
- zur Untersuchung von perzipierten Fremdbildern, etwa in interpersonellen Beziehungen, in Organisationen oder im interkulturellen Vergleich (vgl. Corsini, 1956; Revie, 1956; Shepherd & Guthrie, 1959; Kemnitzer, 1973)
- zur Evaluation klinischer Urteilsbildung (vgl. Rubin & Shontz, 1960; Marks & Seeman, 1962; Graham, 1967; van Atta, 1968).

Faßt man die geschilderten Anwendungsgebiete von Q-Sorts unter dem Aspekt der inhaltlichen Fragestellungen zusammen, die damit untersucht wurden, so wird deutlich, daß Q-Sorts zur Überprüfung sowohl intraindividueller als auch interindividueller Differenzen eingesetzt werden. D.h., es werden sowohl Vergleiche zwischen Individuen als auch Vergleiche innerhalb eines Individuums angestellt. Neben dem Aspekt der Untersuchung inhaltlicher Fragestellungen liegen zahlreiche Studien vor, die sich mit methodologischen Fragen zum Q-Sort auseinandersetzen.

So kommt Frohburg (1970, S. 122) nach einer umfassenden Literatursichtung zu dem Ergebnis, daß das Q-Sort-Verfahren unter methodischem Aspekt zu folgenden Zwecken Verwendung gefunden hat:

- zur Skalierung; etwa für die Beschreibung obengenannter Sachverhalte wie der Charakterisierung von Personen oder Umständen
- zur Interpretation von Ähnlichkeitswerten mehrerer Sortierungen, z.B. für den Vergleich, wie sich eine Person zu verschiedenen Zeitpunkten oder im Vergleich zu anderen Personen darstellt.
- Zur Datengewinnung für faktorenanalytische bzw. varianzanalytische Untersuchungen, etwa zur Fragestellung, ob sich aus den Ergebnissen von Q-Sorts miteinander vergleichbare Persönlichkeitstypen herausfinden lassen.

4. Probleme des Q-Sort-Verfahrens

Vom Grundgedanken her ist das Q-Sort-Verfahren ein Erhebungsinstrument, das theoriegeleitet für den Individualfall konzipiert wird. Konsequenterweise kamen demzufolge in vielen Untersuchungen speziell konstruierte Q-Sorts zur Anwendung. Das wiederum führte zu der Situation, daß dieses Erhebungsverfahren zwar vielfältig benutzt, doch nur unzureichend auf seine Brauchbarkeit hin überprüft wurde. Zudem ließ in vielen Berichten die Beschreibung des eingesetzten Q-Sorts zu wünschen übrig. Demzufolge ist es nahezu unmöglich, vorliegende Ergebnisse zu vergleichen und zu einer vertretbaren Schlußfolgerung hinsichtlich der Angemessenheit der Anwendung dieses Verfahrens zu kommen.

4.1 Internselektion und Itemorganisation

Das Problem der Itemselektion und -Organisation stellt sich insgesamt bei Selbstbeschreibungsverfahren (vgl. Klauer, 1978). Im vorliegenden Fall ist es von besonderer Bedeutung, da die Generalisierbarkeit der Ergebnisse maßgeblich davon beeinflusst wird (vgl. Block, 1961). So konnten Quarter, Kennedy & Laxer (1967) eindrucksvoll belegen, daß die Korrelation zwischen Selbst-

und Idealbild bei adjektivisch formulierten Items höher war als bei solchen, die als Aussagen ausformuliert waren. Des weiteren belegten die Autoren, daß die Reihenfolge der Vorgabe der Sortierung (z. B. erst Selbstbild dann Idealbild vs erst Idealbild dann Selbstbild) keinen Einfluß auf die Höhe der Korrelation zwischen Selbst- und Idealbild haben. In einer solch differenzierten Form wurde sich aber nur selten mit der Problematik auseinandergesetzt.

Im wesentlichen sind drei Itemselektionsmethoden zu unterscheiden, die Block (1961) in Anlehnung an Stephenson differenziert. Die Itemorganisation wird dabei gar nicht besonders bedacht.

Die erste Methode überläßt es dem jeweiligen Untersuchungsleiter, nach subjektiven Kriterien die Bedeutsamkeit von Items für den Untersuchungsgegenstand zu bestimmen. Diesem Ansatz sind die meisten vorliegenden Q-Sort-Verfahren verpflichtet. So konzipierten beispielsweise Butler & Haigh (1954) ihren Q-Sort, indem sie die Items aus den Therapieprotokollen herauszogen, die ihnen zur Verfügung standen. Dabei erfolgte die Zusammenstellung nicht unter einem bestimmten systematischen Gesichtspunkt, sondern sie ergab sich eher willkürlich.

Bei einer zweiten Methode der Itemselektion wird eine operationale Spezifikation des Universums des in Frage stehenden Untersuchungsstands gesucht. So konstruierte Hilden (1958) das 'Universe Of Personal Concepts', indem er jedes Wort, das eine menschliche Reaktion beschreiben kann, aus dem Thorndike Century Senior Dictionary entnahm. Die gefundenen Wörter formulierte er zu kurzen Aussagen. Durch Randomisierung entnahm Hilden aus dem Gesamt der Aussagen 20 kleine Q-Sorts von je 50 Items und verglich die Ergebnisse aus deren Anwendung miteinander. Sowohl die Korrelationen zwischen Selbst- und Idealbild als auch die Differenzierungsfähigkeit zwischen den Personen war zwischen dem Gesamt-Sort und den kleinen Q-Sorts gut vergleichbar. Frohburg (1970, S. 128) kommt aufgrund dieser Ergebnisse zu dem Schluß, „daß die Art des verwendeten Kartensatzes nebensächlich ist, wenn es um die Interpretation von Korrelationen geht“. Ein weiteres Beispiel lieferte Guertin (1973). Er orientierte sich bei der Itemgenerierung an einem bereits erprobten Meßinstrument. Aus Kellys Methode (1955 - personal repertory grid) entwickelte der Autor einen Q-Sort, der dann von ihm faktorenanalytisch untersucht wurde.

Als dritte Methode hatte Stephenson die Itemselektion auf varianzanalytischem Wege vorgeschlagen. Diese Möglichkeit der Konstruktion eines Q-Sorts, die Stephenson (1953) detailliert anhand des Beispiels von Jungs Typologie beschreibt, wurde bis dato weitgehend vernachlässigt. Er ging dabei so vor, daß er aus der Typologie von Jung drei Hauptkonzepte identifizierte: „attitudes“ (Introversion, Extraversion), „mechanisms“ (bewußt, unbewußt) und „functions“ (Denken, Fühlen, Empfinden und Intuition). Die drei Kon-

zepte bilden die unabhängigen Variablen. Es ergibt sich dann ein $2 \times 2 \times 4$ -Design für die Varianzanalyse. Den einzelnen Feldern werden feldspezifische Aussagen zugeordnet. Die so gewonnenen Aussagen werden als Q-Sort-Verfahren den Beurteilern vorgegeben. Die erhaltenen Ergebnisse werden wiederum varianzanalytisch verrechnet. Das grundsätzliche Problem dieses Verfahrens liegt an den fehlenden Kriterien, nach denen die Items ausgewählt werden könnten. Die Willkürlichkeit kann dabei nicht ausgeschlossen werden.

Wesentliche teststatistische Aspekte, die für die Konstruktion eines Q-Sorts und der Selektion von Items bedeutsam sind, wurden in den meisten Studien vernachlässigt. So wurden beispielsweise den Aspekten der Schwierigkeit, Homogenität oder Trennschärfe der Items oder auch der Homogenität der Population der untersuchten Individuen als Basis der Vergleichbarkeit von Ergebnissen wenig Beachtung geschenkt. Das dürfte nicht zuletzt an den testtheoretischen Problemen intraindividueller Messungen liegen. Neff & Cohen (1968) fordern daher, daß bei der Konstruktion eines Sets von Aussagen für einen Q-Sort besonders auf die interne Konsistenz der Items geachtet werden muß. Allerdings sind die Autoren der Ansicht, daß diese Forderung nur für den Fall erfüllt werden kann, daß der Q-Sort auf normativem Weg entwickelt wird. Als Lösung des Problems schlagen die Autoren ein modifiziertes varianzanalytisches Modell vor, in dem ein Koeffizient für die Homogenität der Items innerhalb der Zellen berechnet wird.

4.2 Verteilungsform

Fast ausnahmslos werden die Personen, denen ein Q-Sort-Verfahren vorgelegt wurde, aufgefordert, die vorgegebenen Antwortkategorien mit einer festgelegten Anzahl von Aussagen zu belegen. Damit wird per Instruktion eine wunschgemäße Verteilungsform, meist eine Normal- oder Rechteckverteilung (vgl. Block, 1961), erzielt. Die damit verbundene Annahme, daß sich intraindividuell eingeschätzte Persönlichkeitsmerkmale einer solchen vorgegebenen Verteilungsform entsprechend verteilen, wurde von Jones (1956) geprüft.

Er gab seinen Beurteilern einen ‚free sort‘ vor. Dabei wurden den Personen nur die Kategorien und die Items vorgegeben, ohne nähere Bindung an eine bestimmte Verteilungsform. Die empirisch gefundene Verteilungsform wich signifikant von der einer Normalverteilung ab. Zudem ließ sich auch keine andere Systematik im Sinne einer Ähnlichkeit zu einer bestimmten Verteilungsform erkennen. Damit kann die übliche Instruktion, die zu erzielende Verteilungsform vorzuschreiben, nur methodisch begründet werden. Als Vorteile sind zu werten: die Mittelwerte und Standardabweichungen aller Sortierungen sind gleich, der Fehler der zentralen Tendenz wird reduziert, zudem wird die intraindividuelle Urteilsvariabilität erhöht. Als Nachteile können gel-

ten: die Veränderung der Item-Schwierigkeiten von Anfang bis Ende des Q-Sorts und inhaltliche Verzerrungen aufgrund der Zuordnungsinstruktionen selbst.

4.3 Auswertung

Nach Wylie (1974) sind drei Möglichkeiten der Auswertung von Q-Sort-Daten denkbar, wenn sich die Analyse der Daten auf ein einzelnes untersuchtes Individuum zentriert; in Analogie hierzu wird jedoch auch bei dem interindividuellen Vergleich von Sortierungen verfahren: die Analyse der einzelnen Itemplatzierungen, die Analyse von ‚scores‘ unter einer bestimmten Instruktion (z.B. die Selbstbild-Sortierung) und die Analyse von ‚scores‘ zwischen verschiedenen Instruktionen oder einer Instruktion zu verschiedenen Erhebungszeitpunkten (z.B. Selbstbild- vs Idealbild-Sortierung). Beispiele (vgl. auch oben) für diese drei Verfahrensweisen finden sich bei Dymond (1954). Phillips, Raiford & El-Batrawi (1965) und Frohburg (1972).

Die statistische Auswertung erfolgt über spezielle Computerprogramme (vgl. Baumann, 1970; Wilbur, Gooding & Vincent, 1970). Inhaltlich werden Ähnlichkeitsmaße Verschiedenster Art ermittelt (vgl. Cronbach, 1953; Cohen, 1957; Block, 1961; Coyle, Fowler & Marks, 1967; Frohburg, 1972), die z.T. tabelliert vorliegen und in Rang- oder Produkt-Moment-Korrelationskoeffizienten transformiert werden.

4.4 Gütekriterien

4.4.1 Reliabilität

Insgesamt liegen sehr wenig Untersuchungen vor, die sich mit der Reliabilität von Q-Sorts befaßten (vgl. Steller & Meurer, 1974). Zudem sind kleine Stichprobengrößen und fehlende Angaben zu Zeitintervallen und Repräsentativität der Untersuchungen die Regel (vgl. Dymond, 1954; Frank, 1956; Frohburg, 1972).

Nach Cartwright (1975) sind für das vorliegende Erhebungsinstrument Konsistenz und Stabilität besonders relevant. Mit Konsistenz ist die Frage umschrieben, inwieweit die Items zu verschiedenen Zeitpunkten dasselbe messen; Stabilität meint, inwieweit sich ein Individuum zu verschiedenen Erhebungszeitpunkten als gleich darstellt. Wenn die Reliabilität untersucht wurde, dann unter dem Aspekt der Stabilität. Andere Formen der Reliabilität (z.B. Paralleltest- und Halbierungs-Reliabilität) fehlen ganz (vgl. Livson & Nichols, 1956).

Unter Anlehnung an die Differenzierung von Wylie (1974) zur Auswertung bedeutet die Reliabilität je nach Auswertungsart Unterschiedliches. Auf dem niedrigsten Niveau, der individuellen Itemplatzierung, ließe sich die Frage stellen, inwieweit eine stabile Itemplatzierung von Test zu Test angenommen werden kann. Auf dem nächsten Niveau, Ebene der singulären Sortierung, kann die Frage nach der Zeitstabilität der Rangordnung von Individuen gestellt werden. Auf dem höchsten Niveau, auf der Sorts einer Person unter verschiedenen Instruktionen verglichen werden, ließe sich fragen, ob die Scores stabile intraindividuelle Differenzen widerspiegeln oder nicht.

Die berichteten Stabilitätskoeffizienten sind sehr unterschiedlich (vgl. Dymond, 1954; Frank, 1956; Frohburg, 1972). Das liegt daran, daß unterschiedliche Instrumente benutzt wurden und saubere Vergleichsuntersuchungen fehlen. Eine sorgfältige Studie ist die von Steller & Meurer (1974). Sie ermittelten eine Selbstbildstabilität nach 3 Tagen von $r = .73$ und nach 10 Wochen von $r = .64$. Die entsprechenden Korrelationen der Idealbildstabilität betrugen bei beiden Meßzeitpunkten $r = .74$. Die Selbst-Idealbildstabilitäten waren bei einem Zeitintervall von 3 Tagen $r = .46$ und $r = .59$ und von 10 Wochen $r = .54$ und $r = .57$ (jeweils Anfang und Ende des Zeitintervalls). Die Autoren schlußfolgern daraus: diese Werte 'weisen auf eine gute zeitliche Stabilität dieses Persönlichkeitsmaßes hin' (Steller & Meurer, 1974, S. 621). Ähnliche Ergebnisse bezüglich der Urteilsstabilität wurden gefunden, wenn nicht die eigene Person, sondern Fremdkonzepte Gegenstand der Untersuchung waren (vgl. Burns & Jenkins, 1975).

4.4.2 Validität

Die bisherigen Ausführungen lassen vermuten, daß die Validität von Q-Sorts überwiegend ungeklärt ist. Diese Einschätzung ist zutreffend. Gründe dafür liegen primär darin, daß für spezifische Fragestellungen immer wieder neue Instrumente zusammengestellt werden, ohne den Aufwand ihrer teststatistischen Überprüfung einzugehen. Diese Tendenz wird unterstützt durch den Widerspruch ipsativer Datenerhebung und normativer Datenverrechnung (vgl. Cattell, 1944; Cronbach & Gleser, 1953; Guilford, 1967).

Demzufolge lassen sich zur Validität drei Ansätze differenzieren, die allerdings kein Gesamtbild und damit auch keine Bewertung ermöglichen: die inhaltliche Bestimmung des Q-Sort, das Reflektieren validitätsreduzierender Faktoren beim Q-Sort und Einzelarbeiten zur Überprüfung der Konstruktvalidität eines bestimmten Q-Sorts.

Die inhaltliche Bestimmung des Q-Sorts ist eine Voraussetzung für die Validitätsermittlung. Sie erfolgte theoretisch oder empirisch. Beispiele für ersteres sind der Fremdkonzept-Q-Sort von Block (1961), der Selbstkonzept-Q-Sort

von Butler & Haigh (1954) oder der von Haan und Mitarbeitern zu den Ich-Prozessen „Coping and Defending“ (Haan, 1977).

Beispiele für die empirische inhaltliche Bestimmung verwendeter Q-Sort-Verfahren sind die faktorenanalytischen Studien von Fiedler, 1951; Rogers & Dymond, 1954; Nunnally, 1955; Trush, 1957; Subotnik, 1968; Deo & Hundal, 1969; Woog, 1973.

Bedauerlicherweise wurden die empirisch notwendig sich anschließenden teststatistischen Untersuchungen dann jedoch nicht durchgeführt.

Eine bemerkenswerte Arbeit, in der validitätsreduzierende Faktoren beim Q-Sort-Verfahren insgesamt reflektiert werden, legte Wylie (1974) vor. Zusammenfassend werden die nachfolgenden Faktoren problematisiert:

- Q-Sorts sind nicht frei von validitätsbeeinflussenden Antwortdeterminanten; dazu trägt einerseits die erzwungene Verteilungsform bei, andererseits die mit allen Selbstbeschreibungsinstrumenten verbundene Gefahr, sozial erwünscht zu reagieren (vgl. Taylor, 1955)
- hohe Selbst- und Idealbild-Diskrepanzen sind kontaminiert mit einer allgemeinen Tendenz zu negativen Wahrnehmungen und Beurteilungsprozessen (vgl. Levy, 1956; Kornreich, Straka & Kane, 1968)
- die inhaltliche Bedeutung auftretender Selbst- und Idealbild-Diskrepanzen ist ungeklärt; sie kann für die Beurteiler und den Diagnostiker unterschiedlich sein (vgl. Taylor, 1955)
- die Itemsortierung bei getrennter Vorgabe der Selbst- und Idealbild-Beurteilung ist nicht zwangsläufig dieselbe wie bei einer Sortierung, bei der die zweite Sortierung unter Vorlage der ersten erfolgt
- die korrelative Auswertung impliziert, daß bei vergleichbarer Korrelation Items verschieden sortiert worden sein können.

Diese Probleme sind allerdings nicht hinreichend überprüft und empirisch verfolgt worden. Einzelheiten zur Überprüfung der Konstruktvalidität eines bestimmten Q-Sorts finden sich besonders ausgeprägt im Rahmen der klientenzentrierten Psychotherapieforschung. Überprüft wurde dabei der Butler & Haigh-Q-Sort (1954).

Rogers (1951) ging von der Annahme aus, daß jeder Mensch neben dem Bild, das er von sich hat (Selbstbild) auch eine Vorstellung davon hat, wie er im Idealfall sein möchte (Idealbild). Er nahm weiterhin an, daß gesunde, ausgeglichene, psychisch nicht gestörte Personen eine hohe Übereinstimmung zwischen Selbst- und Idealbild aufweisen (vgl. Sappenfield, 1970). Erfolgreiche Psychotherapie müßte sich demzufolge darin niederschlagen, daß sich eine anfängliche Selbst-Idealbild-Diskrepanz vermindert, indem sich das Selbstbild dem Idealbild annähert.

Diese Annahmen wurden überprüft. Im Vergleich normaler mit psychoneurotischen Klienten bestätigten sich die Vorhersagen (vgl. Hanlon, Hofstätter & O'Connor, 1954; Friedman, 1955; Fagan & Guthrie, 1959; Frohburg, 1972). Die Verwendung des Q-Sort als Therapie-Effekt-Meßinstruments ließ sich für den Neurosebereich stützen (vgl. Rogers & Dymond, 1954; Butler, 1968; Frohburg, 1972; Waskow & Parloff, 1975), nicht jedoch für den Bereich der Psychopathologie (vgl. Ends & Page, 1957; Rogers, 1967).

4.5 Qualität der Daten

Orientiert man sich an einer neueren Systematik zur Differenzierung von Persönlichkeitsdaten, die Block (1977) in Anlehnung an Cattell (1957, 1973) vornimmt, so sind die Daten aus Q-Sort-Verfahren primär S-Daten, d.h., sie werden durch Selbstbeobachtung eigener Verhaltensweisen, Gefühle und Kognitionen gewonnen. Daneben werden jedoch auch Q-Sort-Verfahren zur Erhebung von R-Daten, d.h. Beobachtungsdaten, benutzt. Als Beispiel für diese Form von Q-Sort-Daten steht Block (1961). Die Intention, die dieser Autor mit seinem CQ-Set verfolgt, ist die Einschätzung infragestehender Sachverhalte durch kompetente Beobachter. Als R-Daten lieferndes Instrument gilt der Q-Sort auch dann, wenn statt des Idealbildes externe Standards wie der normale Durchschnittsbürger u.ä. eingeschätzt werden (vgl. Levy, 1956; Fagan & Guthrie, 1959). Diese inhaltliche wohl plausible Verwendung des Q-Sorts weist auf ein zweites zentrales Problem hin. Mit dem Q-Sort-Verfahren wird eine ipsative Messung vorgenommen (Cattell, 1944; Block, 1957; Guilford, 1967). Schon Cattell (1944) wies ausdrücklich darauf hin, daß es unzulässig sei, Daten, die als ipsative Daten gewonnen wurden, wie normative Daten zu behandeln und zu verrechnen. Während Block (1957) nachweisen konnte, daß ipsative und normative Messungen zu vergleichbaren Ergebnissen führen, schränkte Wittenborn (1961) diese Aussagen ein, indem er darauf hinwies, daß die Fehlervarianz bei ipsativen Item-scores, die wie normative behandelt wurden, großen Schwankungen von Item zu Item und von Stichprobe zu Stichprobe unterworfen sein kann. In jüngster Zeit setzte sich Marceil (1977) mit dem Problem idiographischer und nomothetischer Messung erneut auseinander und erarbeitete eine Indikationsmatrix. Diese Matrix beinhaltet einerseits methodische und andererseits theoretische Voraussetzungen der Forschung. Marceil unterteilt diese Voraussetzungen in jeweils zwei weitere Dimensionen.

Dabei subsumiert er unter die methodischen Voraussetzungen die Dimensionen „selektive Untersuchung mehrerer Individuen“ vs „intensive Untersuchung weniger Individuen“, unter die theoretischen Voraussetzungen die Dimensionen „Der Mensch ist eher ähnlich (alike)“ vs „Der Mensch ist eher einzigartig“. Legt man diese Matrix zugrunde und ordnet Fragestellungen den einzelnen Kombinationsmöglichkeiten der Matrixdimensionen zu, so können

in der Forschung idiographische und nomothetische Messungen bei verschiedenen Fragestellungen nebeneinander stehenbleiben.

5. Bedeutung des Q-Sort-Verfahrens

Eine Bewertung des Q-Sort-Verfahrens vorzunehmen ist nicht einfach. Die Schwierigkeiten sind dreifach begründet. Zum einen ist das Verfahren in den 50er Jahren entwickelt worden und zum damaligen Zeitpunkt auch vorrangig eingesetzt und erprobt worden. Das geschah weniger ‚im Sinne des Erfinders‘ als Basis für die Faktorenanalyse einzelner Personen, sondern eher als psychodiagnostisches Instrument zur Messung der intrapsychischen emotionalen Angepaßtheit und Zufriedenheit. Als solches hat das Verfahren auch heute noch seinen Platz behauptet (vgl. Wittenborn, 1961; Waskow & Parloff, 1975). Des weiteren sind Q-Sort-Verfahren vom Konzept her singulär konstruierte Erhebungsinstrumente für spezifische Forschungsfragestellungen. Daraus folgt nahezu zwangsläufig wegen des erheblichen zusätzlichen Forschungsaufwandes, daß die auffindbaren Instrumente zumeist nicht sehr systematisch untersucht worden sind. Letztlich hat man sich in jüngster Zeit mit diesem Erhebungsverfahren nicht mehr gezielt auseinandergesetzt. Das ist um so bedauerlicher, als Testkonstruktion, Testtheorie und Auswertungsmethodik sich seit Begründung des Verfahrens erheblich differenziert haben und das Erhebungsinstrument Q-Sort demzufolge konzeptadäquater als bisher aufgebaut und untersucht werden könnte. Stichwörter in diesem Sinne sind: Interaktionismus, probabilistische Testtheoriemodelle, Einzelfallanalysemethodik, Verfahren zur Bestimmung von Kontentvalidität, Indikation von idiographischen Methoden usw. (Shapiro, 1961; Ekehammar, 1974; Fischer, 1974; Hersen & Barlow, 1976; Marcell, 1977; Klauer, 1978). Eine Neubelebung des differenzierten Einsatzes des Q-Sorts-Verfahrens scheint auch inhaltlich im Rahmen der Ausdifferenzierung der Selbstkonzeptforschung (vgl. Filipp, 1979) und der verstärkt sich entwickelnden klinischen Einzelfalldiagnostik sinnvoll und erfolgversprechend. Ob diese Chance genutzt wird, muß derzeit unbeantwortet bleiben.

Literatur

- Baumann, D. J. 1970. A Computer program for processing Q-sort data. *Educational and Psychological Measurement*, 30, 167-168.
- Block, J. 1957. A comparison between ipsative and normative ratings of personality. *Journal of Abnormal Social Psychology*, 54, 50-54.
- Block, J. 1961. *The Q-sort Method in personality assessment and psychiatric research*. Springfield, Ill.: Charles C. Thomas.

- Block, J. 1977. Advancing the psychology of personality: Paradigmatic shift or improving the quality of research? In: D. Magnusson & N. S. Endler (eds): *Personality at the crossroads: Current issues in interactional psychology*. New York: Wiley & Sons.
- Burns, E. & Jenkins, E. 1975. Stability of Q-sorts in assessing descriptions of hyperactivity. *Perceptual and Motor Skills*, 40, 694.
- Butler, J. M. 1968. Self-ideal congruence in psychotherapy. *Psychotherapy: Theory, Research, and Practice*, 5, 13-17.
- Butler, J. M. & Haigh, G. V. 1954. Changes in the relation between self-concepts and ideal-concepts. In: C. R. Rogers & R. F. Dymond (eds): *Psychotherapy and personality change*. Chicago: University of Chicago Press.
- Caplan, S. W. 1957. The effect of group counseling on junior high school boy's concepts of themselves in school. *Journal of Counseling Psychology*, 4, 124-128.
- Cartwright, D. S. 1975. Patient self-report measures. In: I. E. Waskow & M. B. Parloff (eds): *Psychotherapy change measures*. Washington: U. S. Government Printing Office.
- Cartwright, R. D. 1957. Effects of psychotherapy on self-consistency. *Journal of Counseling Psychology*, 4, 15-29.
- Cattell, R. B. 1944. Psychological measurement: Ipsative, normative, and interactive. *Psychological Reviews*, 51, 292-303.
- Cattell, R. B. 1957. *Personality and motivation structure and measurement*. Yonkers-on-Hudson, New York: World Book Company.
- Cattell, R. B. 1973. *Personality and mood by questionnaire*. San Francisco: Jossey-Bass.
- Chase, P. H. 1957. Self concepts in adjusted and maladjusted hospital patients. *Journal of Consulting Psychology*, 21, 495-497.
- Cohen, J. 1957. An aid in the computation of correlations based on Q-sorts. *Psychological Bulletin*, 54, 138-139.
- Corsini, R. J. 1956. Understanding and similarity in marriage. *Journal of Abnormal Social Psychology*, 52, 327-332.
- Coyle, F. A. Fowler, R. D. & Marks, P. A. 1967. Methodological aide in correlating personality descriptions using the Marks Q-sort. *Psychological Reports*, 21, 563-564.
- Cronbach, L. J. 1953. Correlations between persons as a research tool. In: O. H. Mowrer (ed.): *Psychotherapy: Theory and research*. New York: Ronald Press.
- Cronbach, L. J. & Gleser, G. C. 1953. Assessing similarity between profiles. *Psychological Bulletin*, 50, 456-473.
- Deo, P. & Hundal, B. S. 1969. Self-concept types by Q-Technique. *Journal of Psychological Researches*, 13, 1-11.
- Dorsch, F. 1976⁹. *Psychologisches Wörterbuch*. Bern: Huber.
- Dymond, R. F. 1954. Adjustment changes over therapy from self-sorts. In: C. R.

- Rogers & R. F. Dymond (eds): *Psychotherapy and personality change*. Chicago: University of Chicago Press.
- Ekehammar, B. 1974. Interactionism in personality from a historical perspective. *Psychological Bulletin*, 81, 1026-1048.
- Ends, E. J. & Page, C. W. 1957. A study of three types of group psychotherapy with hospitalized male inebriates. *Quarterly Journal of Studies on Alcohol*, 18, 263-277.
- Engel, M. 1959. The stability of the self-concept in adolescence. *Journal of Abnormal Psychology*, 58, 211-215.
- Fagan, J. & Guthrie, G. M. 1959. Perception of self and of normality in schizophrenics. *Journal of Clinical Psychology*, 15, 203-207.
- Fiedler, F. E. 1951. Factor analyses of psychoanalytic, nondirective, and Adlerian therapeutic relationships. *Journal of Consulting Psychology*, 23, 177-180.
- Filipp, S. H. 1979. *Selbstkonzept-Forschung: Probleme, Befunde, Perspektiven*. Stuttgart: Klett-Cotta.
- Fischer, G. H. 1974. *Einführung in die Theorie psychologischer Tests: Grundlagen und Anwendungen*. Bern: Huber.
- Frank, G. H. 1956. Note on the reliability of Q-sort data. *Psychological Reports*, 2, 182.
- Friedman, I. 1955. Phenomenal, ideal, and projected conceptions of self. *Journal of Abnormal and Social Psychology*, 51, 611-615.
- Frohburg, I. 1972. Die Verwendbarkeit psychodiagnostischer Methoden zur Veränderungsmessung in der Psychotherapie. In: J. Helm (ed.): *Psychotherapieforschung: Fragen, Versuche, Fakten*. Berlin: VEB.
- Frohburg, I. 1970. Zur psychodiagnostischen Erfassung von Persönlichkeitsveränderungen mit Hilfe der Q-Sortierungstechnik. In: H. D. Rösler, H. D. Schmidt & H. Szweczyk (eds): *Persönlichkeitsdiagnostik*. Berlin: VEB
- Garfield, S. L. & Prager, R. A. 1971. Evaluation of outcome in psychotherapy. *Journal of Consulting and Clinical Psychology*, 37, 307-313.
- Graham, J. R. 1967. A Q-sort study of the accuracy of clinical descriptions based on the MMPI. *Journal of Psychiatric Research*, 5, 297-305.
- Guertin, W. H. 1973. Sorto: Factor analyzing Q-sorts of Kellys personal construct productions. *Journal of Personality Assessment*, 37, 69-77.
- Guilford, J. P. 1967. When not to Factor Analyse. In: D. N. Jackson & S. Messick (eds): *Problems in human assessment*. New York: McGraw-Hill.
- Haan, N. 1977. *Coping and defending: Processes of self-environment Organisation*. New York: Academic Press.
- Hanlon, R. E., Hofstaetter, P. & O'Connor, J. 1954. Congruence of self and ideal self in relation to personality adjustment. *Journal of Consulting Psychology*, 18, 215-218.
- Hartley, M. W. 1950. *Q-Technique: Its methodology and application* (unveröffentl.).

- Hersen, M. & Barlow, D. H. 1976. Single case experimental designs: Strategies for studying behavior change in the individual. New York: Pergamon Press.
- Hilden, A. H. 1958. Q-sort correlation: Stability and random choice of statements. *Journal of Consulting Psychology*, 22, 45-50.
- John, D. & Keil, W. 1972. Selbsteinschätzung und Verhaltensbeurteilung. *Psychologische Rundschau*, 23, 10-29.
- Jones, A. 1956. Distributions of traits in current Q-sort methodology. *Journal of Abnormal Social Psychology*, 53, 90-95.
- Kalis, B. L. & Bennett, L. F. 1957. The assessment of communication: The relation of clinical improvement to measured changes in communicative behavior. *Journal of Consulting Psychology*, 21, 10-14.
- Kelly, G. A. 1955. The psychology of personal constructs. New York: Norton.
- Kemnitzer, L. S. 1973. Adjustment and value conflict in urbanizing Dakota Indians measured by Q-sort technique. *American Anthropologist*, 75, 687-707.
- Kerlinger, F. N. 1956. The attitude structure of the individual: A Q-Study of the educational attitudes of Professors and Laymen. *Genetic Psychology Monographs*, 53, 283-329.
- Klauer, K. J. 1978. Handbuch der Pädagogischen Diagnostik. Düsseldorf: Pädagogischer Verlag Schwann.
- Kornreich, b., Straka, J. & Kane, A. 1968. Meaning of self-image disparity as measured by the Q-Sort. *Journal of Consulting and Clinical Psychology*, 32, 728-730.
- Langer, J. & Schulz v. Thun, F. 1974. Messung komplexer Merkmale in der Psychologie und Pädagogik. München: Reinhardt.
- Levy, L. H. 1956. The meaning and generality of perceived actual-ideal discrepancies. *Journal of Consulting Psychology*, 20, 396-398.
- Livson, N. H. & Nichols, T. F. 1956. Discrimination and reliability in Q-sort personality descriptions. *Journal of Abnormal and Social Psychology*, 52, 159-165.
- Marcill, J. C. 1977. Implicit dimensions of idiography and nomothesis: A reformulation. *American Psychologist*, 32, 1046-1055.
- Marks, P. A. & Seeman, W. 1962. The heterogeneity of some common psychiatric stereotypes. *Journal of Clinical Psychology*, 18, 266-270.
- Minsal, W.-R. & Bente, G. 1979. Gesprächspsychotherapie. In: W. Wittling (ed.): Handbuch der klinischen Psychologie, Bd. 1: Methoden der klinischen Psychologie. Hamburg: Hoffmann & Campe.
- Mowrer, H. H. 1953. Q-technique - description, history and critique. In: O. H. Mowrer (ed.): Psychotherapy theory and research, 1953. New York: Ronald Press.
- Nahisnky, I. D. 1958. The relationship between the self-concept and the ideal-self concept as a measure of adjustment. *Journal of Clinical Psychology*, 14, 360-364.
- Neff, W. S. & Cohen, J. 1967. A method for the analysis of the structure and internal consistency of Q-sort arrays. *Psychological Bulletin*, 68, 361-368.

- Nunnally, J. C. 1955. A systematic approach to the construction of hypothesis about the process of psychotherapy. *Journal of Consulting Psychology*, 19, 17-20.
- Perkins, H. V. 1958. Factors influencing change in children's self-concepts. *Child Development*, 29, 221-230.
- Peterson, A. O. D., Snyder, W. U., Guthrie, M. & Ray, W. S. 1958. Therapist factors: An exploratory investigation of therapeutic biases. *Journal of Counseling Psychology*, 5, 169-173.
- Phillips, E. L., Raiford, A. & El-Batrawi, S. 1965. The Q-sort reevaluated. *Journal of Consulting Psychology*, 29, 422-425.
- Quarter, J., Kennedy, D. R. & Laxer, R. M. 1967. Effect of order and form in the Q-sort. *Psychological Reports*, 20, 893-894.
- Revie, V. A. 1956. The effect of psychological case work on the teacher's concept of the pupil. *Journal of Counseling Psychology*, 3, 125-129.
- Rogers, A. H. 1958. The self-concept in paranoid schizophrenia. *Journal of Clinical Psychology*, 14, 365-366.
- Rogers, C. R. 1951. *Client-centered therapy*. Boston: Houghton Mifflin.
- Rogers, C. R. 1967. *The therapeutic relationship and its impact: A study of psychotherapy and schizophrenics*. Madison: University of Wisconsin Press.
- Rogers, C. R. & Dymond, R. F. 1954. *Psychotherapy and personality change*. Chicago: University of Chicago Press.
- Rubin, M. & Shontz, F. C. 1960. Diagnostic prototypes and diagnostic process of clinical psychologists. *Journal of Consulting Psychology*, 24, 234-239.
- Sappenfield, B. R. 1970. Perception of self as related to perception of the „ideal personality“. *Perceptual and Motor Skills*, 31, 975-978.
- Schön, G.-H. 1966. Die Anwendung des „Q-Sort-Verfahrens“ zur Quantifizierung von Persönlichkeitsbeurteilungen in der klinischen Psychologie. Psychologisches Institut Hamburg (unveröffentl.).
- Shapiro, M. B. 1961. A method of measuring psychological changes specific to the individual psychiatric patient. *British Journal of Medical Psychology*, 34, 151-155.
- Shepherd, I. L. & Guthrie, G. M. 1959. Attitudes of mothers of schizophrenic patients. *Journal of Clinical Psychology*, 15, 212-215.
- Sherry, P. & Hurley, J. R. 1976. Curative factors in psychotherapeutic and growth groups. *Journal of Clinical Psychology*, 32, 835-837.
- Shlien, J. M. 1964. Comparison of results with different forms of psychotherapy. *American Journal of Psychotherapy*, 18, 15-22.
- Shontz, F. C. 1956. Evaluative conceptualisations as the basis for clinical judgements. *Journal of Consulting Psychology*, 20, 212-215.
- Steller, M. & Meurer, K. 1974. Zur Reliabilität eines Q-Sort zur Veränderungsmessung. *Psychologische Beiträge*, 16, 618-624.

- Stephenson, W. 1953. The study of behavior. Chicago: University of Chicago Press.
- Subotnik, L. 1968. Transference in Client-Centered Child Therapy. An unsuccessful case. *Journal of Genetic Psychology*, 112, 183-189.
- Taylor, D. M. 1955. Changes in the self concept without psychotherapy. *Journal of Consulting Psychology*, 19, 205-209.
- Tobacyk, J. J., Broughton, A. & Vaught, G. M. 1975. Effects of congruence - incongruence between locus of control and field dependence on personality functioning. *Journal of Consulting and Clinical Psychology*, 43, 81-85.
- Trush, R. S. 1957. An agency in transition: The case study of a counseling center. *Journal of Counseling Psychology*, 4, 183-190.
- Turner, R. H. & Vanderlippe, R. H. 1958. Self-ideal congruence as an index of adjustment. *Journal of Abnormal Social Psychology*, 57, 202-206.
- Van Atta, R. E. 1968. Concepts employed by accurate and inaccurate Clinicians. *Journal of Counseling Psychology*, 15, 338-345.
- Walker, R. N. 1968. A scale for parent's ratings: some ipsative and normative correlations. *Genetic Psychology Monographs*, 77, 95-133.
- Waskow, J. E. & Parloff, M. B. 1975. Psychotherapy Change measures. Washington: U. S. Government Printing Office.
- Wilbur, P. H., Gooding, C. T. & Vincent, R. A. 1970. Adapting Q-technique to Computer scoring procedures. *Educational and Psychological Measurement*, 30, 169-170.
- Wittenborn, J. R. 1961. Contributions and current status of Q-Methodology. *Psychological Bulletin*, 58, 132-142.
- Woog, P. C. 1973. A Q-study of elementary school teachers assignments of educational priorities and their practice. *Journal of Experimental Education*, 42, 88-96.
- Wylie, R. C. 1974². The self-concept: A review of methodological considerations and measuring instruments. Lincoln: University of Nebraska Press.

4. Kapitel

Semantische Differential Technik

Bernd Schäfer

1. Einleitung

Die Semantische Differential Technik ist eine Methode zur Analyse der Bedeutung von Zeichen. Ein Semantisches Differential (SD) besteht aus einer (nicht verbindlich festgelegten) Anzahl von bipolaren (meist siebenstufigen) Rating-skalen, deren Endpunkte in der Regel durch Adjektive gekennzeichnet sind. Im deutschen Sprachraum werden u.a. auch die Bezeichnungen ‚Eindrucksdifferential‘ und ‚Polaritätsprofil‘ verwendet. Die SD-Technik wurde von Osgood und Mitarbeitern (Osgood 1952; Osgood & Suci 1955; Osgood et al. 1957) zur Analyse der dem sprachlichen Bedeutungsverhalten zugrunde liegenden Dimensionalität entwickelt und wird seitdem auch außerhalb der psycholinguistischen Problemstellung mit großer Häufigkeit in der empirischen Sozialforschung und in nahezu allen Bereichen der psychologischen Forschung eingesetzt. Darstellungen theoretischer, methodischer und technischer Art geben Osgood, Suci & Tannenbaum (1957), Heise (1969), Snider & Osgood (1969), Bergler (1975) und Osgood, May & Miron (1975).

1.1 Zugrundeliegende Modelle

Die SD-Technik wird von Hörmann (1976, 92) im Hinblick auf die Komplexität ihrer theoretischen Begründung als „eine glänzende Leistung des ‚aufgeklärten Neobehaviorismus‘“ gewürdigt. Als Methode zur Erfassung bedeutungsspezifischer Reaktionen auf Zeichen ist sie durch eine Verhaltenstheorie der Bedeutung von Zeichen fundiert, die mit einem Mess- und einem Raummodell in einen Korrespondenz-Zusammenhang eingebettet ist. Zwar sind diese drei Modelle logisch voneinander unabhängig, ihre Verbindung kann aber als charakteristisch für die SD-Technik gelten.

1.1.1 Verhaltensmodell (*representational mediation theory*)

Die Bedeutung von Zeichen wird vom beobachtbaren Verhalten gegenüber den bezeichneten Dingen hergeleitet. Die Genese von Zeichen und ihrer Bedeutung wird von Osgood (1971, 11) folgendermaßen beschrieben:

„a stimulus pattern (S') which is not the same physical event as the thing signified (S) will become a sign of that significate when it becomes conditioned to a mediation process, this process: (a) being some distinctive representation of the total behavior (R_T) produced by the significate, and (b) serving to mediate overt behaviors (R_X) to the sign which are appropriate to („take account of“) the significate.



Abb. 1: (Nach Osgood et al. 1957, 7)

a) Entwicklung primärer Zeichen

b) Entwicklung sekundärer Zeichen

Das potentielle Zeichen S' , z.B. ein Wort, löst also nach raum-zeitlicher Verbindung mit einem Ding S einen (reduzierten) Teil des Gesamtverhaltens auf S aus, der - bei alleiniger Darbietung des Zeichens - als repräsentationale Response r_M die Funktion hat, ein Autostimulationsmuster s_M zu vermitteln, das die zeichenspezifischen, dem bezeichneten Sachverhalt S rechnungtragenden, offenen Verhaltensweisen R_X auslöst. Die Bedeutung eines Zeichens ist nach dieser Konzeption durch einen spezifischen repräsentationalen Vermittlungsprozeß bestimmt.

Die Mehrzahl aller Zeichen hat ihre Bedeutung durch Verbindung mit anderen Zeichen und nicht unmittelbar mit den bezeichneten Dingen erhalten (sekundäre Zeichen, vgl. Abb. 1b). Die meisten Sechsjährigen, von denen die wenigsten jemals einem Zebra begegnet sind, verstehen das Wort ‚Zebra‘: sie haben Bilder von Zebras gesehen, gehört, daß Zebras gestreift sind, wie Pferde laufen und gewöhnlich wild leben (Osgood et al. 1957, 8). Das Reizmuster ‚Zebra‘ (S'') erhält Teile derjenigen Mediations-Responses r_M , die bereits von den primären Zeichen ausgelöst werden.

Für die bedeutungsspezifischen r_M gilt, daß sie aus Komponenten (r_i) bestehen; die Eigenart der r_M wird jeweils durch die spezifische Kombination der r_i bestimmt.

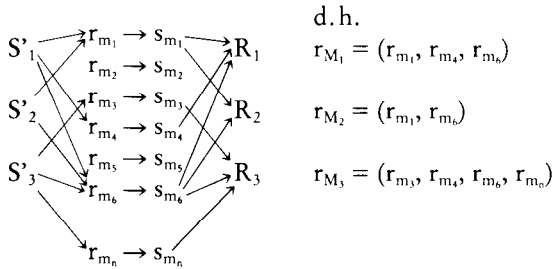


Abb. 2: (Nach Osgood 1971, 12) Komponenten zeichenspezifischer Vermittlungsprozesse.

Eine Diskussion der Osgoodschen Bedeutungskonzeption findet sich bei Fuchs (1975a).

1.1.2 Meßmodell

Um die Bedeutung von Zeichen zu erfassen, müssen der Mediationskonzeption gemäß die repräsentationalen Vermittlungsprozesse ($r_M \rightarrow s_M$) zugänglich gemacht werden. Diese äußern sich, abhängig von der Stärke der Verbindung mit den Autostimulationsprozessen und situativen Bedingungen, in den offenen Verhaltensweisen R_X . Die von Osgood und Mitarbeitern vorgeschlagene Lösung ist die SD-Technik: eine Kollektion von siebenstufigen bipolaren Ratingskalen, deren Merkmale eine repräsentative Stichprobe bedeutungsspezifischen Urteilsverhaltens darstellen und dazu dienen sollen, Zeichen quantitativ zu qualifizieren, z.B. Hans ist (außerordentlich) klug, (sehr) mächtig, (etwas) alt usw.

HANS									
dumm	—	:	—	:	—	:	—	:	x
mächtig	—	:	x	:	—	:	—	:	—
jung	—	:	—	:	—	:	x	:	—
<hr/>									
	-3		-2		-1		0		1
									2
									3

1.1.3 Raummodell

In der Analogie einer räumlichen Darstellung von Bedeutung lassen sich Zeichen als Punkte in einem geometrischen Raum lokalisieren. Die Bedeutung eines Zeichens wird durch einen Vektor repräsentiert, der vom Schnittpunkt der Achsen, dem Punkt vollständiger Bedeutungslosigkeit, ausgeht: die Länge des Vektors entspricht der Bedeutsamkeit (Intensität, Sättigung), seine Rich-

tung der ‚semantischen Qualität‘ des Zeichens. Die Bedeutungsähnlichkeit von Zeichen wird durch die Größe der Distanzen zwischen den Punkten abgebildet.

Als Achsen eines derartigen Bedeutungsraumes könnten die SD-Ratingskalen aufgefaßt werden. Da sie einen Raum konstituieren würden, dessen Ordnung beliebig bliebe, wird die empirische Analyse der dimensionalen Struktur zum zentralen Problem der Entwicklung des Bedeutungsraumes.

1.2 Integration der Modelle

Osgood (1971) legt Wert auf die Darlegung, daß diese drei Modelle streng aufeinander bezogen sind.

1. Die Komponenten r_m der repräsentationalen Responses r_M werden mit den empirisch gewonnenen Hauptachsen des Bedeutungsraumes, den Dimensionen E(valuation), P(otency) und A(ctivity) identifiziert. In ihrer Verschiedenheit repräsentieren die Bedeutungs-Komponenten E, P und A nach Osgoods Auffassung solche Aspekte des Verhaltens gegenüber Dingen, die aufgrund unterschiedlicher Anpassungsfunktion differentiell verstärkt werden. In dieser Betrachtungsweise sind Zeichen seit den Zeiten des Neandertalers dadurch bedeutsam, d.h. verhaltensrelevant, daß sie vor allem spezifizieren, in welchem Maße die bezeichneten Dinge ‚gut‘ oder ‚schlecht‘ (E), ‚stark‘ oder ‚schwach‘ (P), ‚aktiv‘ oder ‚passiv‘ (A) sind.
2. Da die offenen Reaktionsweisen, die von den Vermittlungsprozessen repräsentiert werden, nach einem reziprok-antagonistischen Muster organisiert seien, folge, daß die Komponenten r_m in eben dieser Weise funktionierten. Da die offenen Reaktionsweisen im Hinblick auf ihre Intensität variieren, wird auch ihren repräsentationalen Vermittlungsprozessen diese Eigenschaft zugeschrieben. Aufgrund der Unvereinbarkeit gleichzeitiger Tendenzen in Richtung auf antagonistische r_m (also Tendenzen z.B. in Richtung E+ und E-) wird angenommen, daß sich beide, wenn sie auftreten, in Richtung auf Neutralität oder Bedeutungslosigkeit aufheben.

Dem reziprok-antagonistischen Charakter der bedeutungsspezifischen Mediatoren und ihrer Intensitätsvariation wird durch die Verwendung bipolarer, quantitativ abgestufter Beurteilungsskalen Rechnung getragen.

3. Schließlich werden die Punkte im Raum, die die Bedeutung von Zeichen repräsentieren, mit den r_M insgesamt, wie sie durch Zeichen hervorgerufen werden, identifiziert.

Osgood bezeichnet die Beziehung der in Abb. 3 wiedergegebenen Modelle als „isomorph“.

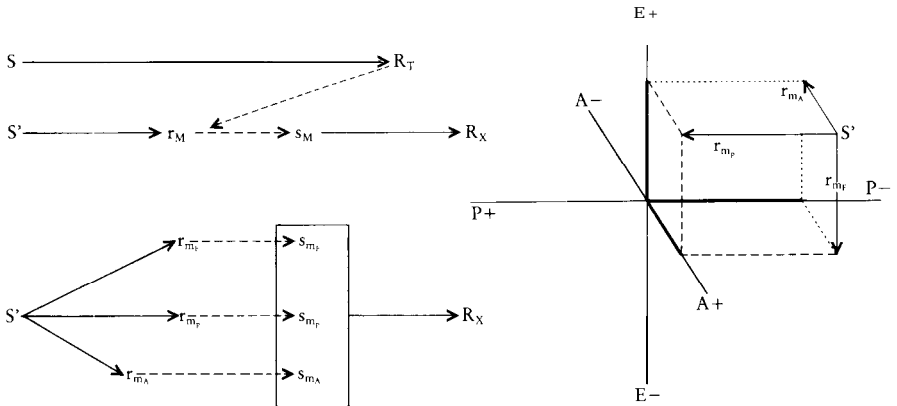


Abb. 3: (Nach Osgood 1962, 11 und 1971, 17) Integration der Modelle („Isomorphismus“)

2. Ordnung von SD-Daten: Architektur eines universellen Bedeutungsraumes

Die Bedeutung eines Zeichens ist - der Konzeption Osgoods gemäß - durch seine Lage im Bedeutungsraum bestimmt. Das zentrale Problem der SD-Forschung betrifft die Identifizierung der Dimensionalität dieses Raumes; die dazu vorliegende Lösung ist die EPA-Struktur. Aufgrund von Belegen ihrer Stabilität gegenüber der Variation von Beurteilern und Konzepten (sprachlichen Zeichen, meist Nomen) und zwar sowohl innerhalb wie zwischen Sprachen/Kulturen gilt sie als ein generelles bzw. universelles Sprachmerkmal. Der anwendungsorientierten Forschung dient sie als ein allgemeines Bezugssystem für quantitative Vergleiche der Bedeutung von Zeichen.

Es verdient an dieser Stelle hervorgehoben zu werden, daß die SD-Technik nur in Verbindung mit dem Modell eines Semantischen Raumes als eine eigenständige Forschungstechnik anzusehen ist. Abgelöst vom Raummodell stellt ein SD lediglich eine Ansammlung von Ratingskalen dar. Allerdings ist die SD-Technik keineswegs an die EPA-Lösung der Struktur von Zeichen-Bedeutung gebunden. Im folgenden sollen zunächst die grundlegenden Befunde zur Geltung dieses semantischen Faktoren-Systems skizziert, sowie Bedingungen erörtert werden, die das Auftreten der EPA-Struktur beeinflussen.

Wichtigste Erkenntnisgrundlage sind die Forschungsberichte ‚The Measurement of Meaning‘ (Osgood et al. 1957) und ‚Cross-Cultural Universals of Affective Meaning‘ (Osgood et al. 1975), in denen ein beispiellos ehrgeiziges, mehrphasiges semantisches ‚Raumfahrt‘-Programm beschrieben wird, das Osgood und Mitarbeiter zusammen mit ‚Copiloten‘, computerorientiertem ‚Bodenpersonal‘ und ‚Beobachtungsstationen‘ rund

um die Erde durchgeführt haben. Nur die kritische Würdigung der Befunde zur Konzeption des Semantischen Raumes trägt zur Klärung der praktisch bedeutsamen Frage bei, inwieweit das SD als eine wiederverwendungsfähige semantische ‚Raumfähre‘ eingesetzt werden kann und wann die Exploration semantischer Räume spezifischer SDs bedarf.

Die folgende Darstellung geht von den für SD-Erhebungen typischen Daten aus. SD-Urteile werden von Beurteilern auf Skalen für Konzepte abgegeben; die Daten lassen sich also in einer dreimodalen Matrix anordnen (vgl. Abb.4).

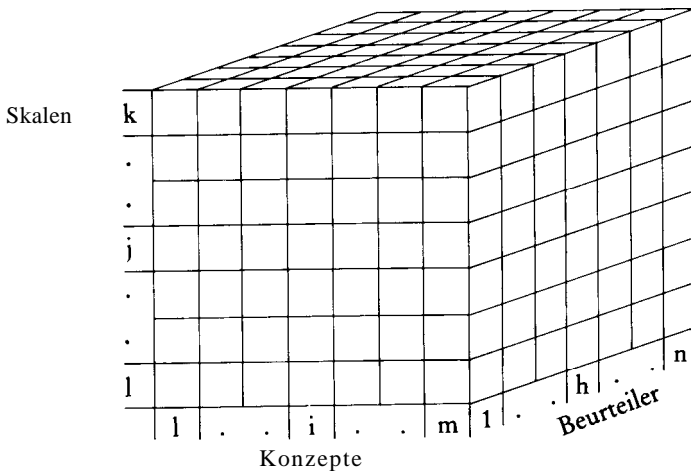


Abb. 4: Dreimodaler SD-Datenkubus

2.1 Skalen-Kovariation: Generalität der EPA-Struktur

2.1.1 Grundlegende Befunde (*The Measurement of Meaning: Osgood et al. 1957*)

In einer ersten Analyse wurden zu 40 Nomen (der Kent-Rosanoff-Liste) bei Studenten Primärassoziationen (Adjektive) gesammelt. Die fünfzig häufigsten wurden zusammen mit ihren Antonymen in der Standardform 100 Studenten zur Beurteilung von 20 (anderen) Konzepten vorgelegt. Die Skalenkorrelationen (über Beurteiler und Konzepte) wurden nach der Centroid-Methode faktorisiert und orthogonal rotiert. Als Ergebnis wurde die klassische EPA-Lösung gefunden: Evaluation (33,8 % ges. Varianz) mit hohen Ladungen auf den Skalen ‚good - bad‘, ‚beautiful - ugly‘, ‚sweet - sour‘; Potency (7,6% ges. Varianz) durch ‚large - small‘, ‚strong - weak‘, ‚heavy - light‘; Activity (6,2% ges. Varianz) mit ‚fast - slow‘, ‚active - passive‘ und ‚hot - cold‘.

Um die Unabhängigkeit dieser Lösung von der Auswahl der Konzepte zu überprüfen, wurden in einer zweiten Analyse alle fünfzig Adjektivpaare einer Stichprobe von 40 Beurteilern aus der gleichen Population paarweise vorgelegt. Ein Adjektiv des ersten Paares diente als ‚Konzept‘; das zu ihm am besten passende Adjektiv des anderen Paares sollte ausgewählt werden. Die Prozentwerte der Wahlhäufigkeiten in einer Skalen x Skalen-Matrix wurden in einer für Rohdaten modifizierten Form der Diagonalmethode nach Thurstone faktorisiert. Die Ähnlichkeitsbeziehungen der Skalen, wie sie sich in dieser Prozentwerte-Matrix ausdrücken, setzen jeweils perfekte Bipolarität voraus - eine Annahme, die angesichts der ad-hoc-Erzeugung der Antonyme und der inzwischen vorliegenden Befunde (vgl. Abschn. 3.1) problematisch erscheint. Die Ergebnisse wiesen befriedigende Übereinstimmung mit denen der ersten Analyse auf; die Autoren konstatieren, daß die EPA-Struktur damit gegenüber einem Wechsel der Beurteilerstichprobe und der Datenerhebungsmethode stabil geblieben ist.

Als Beitrag zur Begründung der Generalität der EPA-Struktur werden in den frühen Arbeiten des Measurement of Meaning die Ergebnisse einer dritten Studie interpretiert. Um die Abhängigkeit dieser Lösung von der vorher getroffenen Skalenauswahl einzuschätzen und mögliche weitere Dimensionen des Semantischen Raumes zu identifizieren, wurden die Beschreibungsmerkmale nun systematisch aus Roget's Thesaurus ausgewählt. Wiederum 100 Studenten beurteilten mit Hilfe von 76 Skalen (Kapazitätsgrenze des Computers) 2.0 Konzepte unterschiedlicher Kategorienzugehörigkeit. Da die unrotierte Centroid-Lösung die EPA-Struktur erkennen ließ, wurde sie durch Pivot-Skalen fixiert; für die Residualvarianz wurden fünf weitere Faktoren tentativ interpretiert. Der erheblich geringere Anteil der durch EPA erklärten Varianz (28 % gegenüber 48 % bzw. 44 % der ges. Varianz), der im wesentlichen zu Lasten des E-Faktors ging, wird von den Autoren auf die Art des Skalensamplings zurückgeführt.

Ergänzende Analysen, die die EPA-Struktur auch für die Beurteilung von nicht-verbalen Konzepten (Ortungssignale, Gemälde) belegen, werden berichtet.

Die frühen Arbeiten deuten auf eine spezifische, durch EPA beschriebene Struktur des Urteilsverhaltens hin, die gegenüber einer gewissen Variation der Datenanalysetechnik (im Bereich der Faktorenanalyse), dem Kriterium der Auswahl ‚repräsentativer‘ Skalen, der Auswahl von Konzepten und der Wahl der Stichprobe von Beurteilern aus einer Population (Studenten) zumindest qualitativ stabil bleibt.

Eine systematische Untersuchung der EPA-Generalität hat die Art des Samplings in allen drei Varianzquellen, sowie die Aufklärung der gesamten, in den Daten enthaltenen Varianz zu berücksichtigen.

2.1.2 Berücksichtigung der verfügbaren Varianz von SD-Daten

Die den drei Modalitäten entsprechenden Varianzquellen sind für die Analyse der Dimensionalität des Semantischen Raumes in unterschiedlicher Weise relevant: Der Semantische Raum wird durch Achsen bestimmt, deren Definition auf der Kovariation zwischen den Skalen basiert. Gesucht wird die geringste Anzahl von Achsen, die Komponenten r_m der r_M des mediationstheoretischen Modells, die das bedeutungsspezifische Verhalten hinreichend beschreiben. Für die Autoren des Measurement of Meaning (Osgood et al. 1957), wie die Autoren von Cross-Cultural Universals of Affective Meaning (Osgood et al. 1975) ist die Faktorenanalyse die Methode der Wahl zur Lösung dieses Problems gewesen und geblieben.

2.1.2.1 Daten-Reduktionstechniken

Die während der frühen Arbeiten verfügbaren faktorenanalytischen Modelle waren nur für zweimodale Datenmatrizen anwendbar. Für die Überführung der dreimodalen Ausgangsmatrix in eine Korrelationsmatrix diskutieren Miron & Osgood (1966) drei Reduktionstechniken:

„stringing out“

Bei dieser, häufig und auch von Osgood et al. (1957) verwendeten Technik werden die Korrelationen zwischen den Skalen für die ‚aufgereihten‘ Markierungen aller Beurteiler bei allen Konzepten berechnet. Miron & Osgood (1966; vgl. auch Osgood et al. 1975, 45f und insbes. Miron 1972, 315ff) bemängeln, daß hier die Struktur der Skalenvariation nicht nur im Hinblick auf die Konzepte, sondern auch gleichzeitig durch Beurteilervarianz bestimmt wird. Bei geringer Konzeptvarianz determiniere diese u.U. entscheidend die Skalenvariation.

Summation

Die Beurteilervarianz wird durch Summierung über die Beurteiler innerhalb der Konzepte reduziert. Die Skalen-Korrelationen werden für die Summenwerte bei den Konzepten bestimmt. Diese Methode wurde im Rahmen des kulturvergleichenden Projekts (Osgood et al. 1975) bevorzugt.

Durchschnittliche Korrelation

Die für alle Beurteiler gesondert berechneten Skalen-Korrelationsmatrizen (für alle Konzepte) werden über die Beurteiler gemittelt.

Während also im ersten Falle die gesamte Beurteiler- und Konzeptvarianz bei der Berechnung der durchschnittlichen Skalenvariationen berücksichtigt wird, wird die Beurteilervarianz in den beiden anderen Verfahren als Fehlervarianz betrachtet und durch Mittelung reduziert. Insbesondere ein der ‚Summation‘ entsprechendes Vorgehen muß schon deshalb empirisch gut begründet werden, weil es - statt einer strengen Prüfung - die generelle (d.h. u.a. auch beurteiler-unabhängige) Geltung der EPA-Struktur als eine Folge der metho-

dischen Manipulation implizieren kann: Beurteilervarianz wird hier gänzlich ignoriert. Dafür, daß sie zu Recht vernachlässigbar ist, werden folgende Argumente vorgebracht (Osgood et al. 1957; Miron & Osgood 1966; Osgood et al. 1975):

- Faktorenanalysen aufgrund von Skalengkorationen nach der ‚stringing-out‘ und der Summations-Methode führen (in einem Beispielfall) zu sehr ähnlichen Ergebnissen.
- Faktorenanalysen von Beurteiler x Beurteiler-Matrizen mit den Korrelationen der beurteilerspezifischen Skalengkorationen führen (in zwei Referenzfällen: Ware 1958; Tanaka & Osgood 1965) zu Lösungen, die nur einen Beurteilerfaktor bedeutsam erscheinen lassen.
- Die Ergebnisse von seit neuerem vorliegenden dreimodalen Faktorenanalysen lassen erkennen, daß der ‚Löwenanteil‘ der Beurteilervarianz jeweils durch nur einen Faktor erklärt wird.

Die Analyse interindividueller Differenzen wird in Abschn. 2.1.5 noch ausführlicher erörtert.

2.1.2.2 Konzeptvarianz

Während die Autoren des Measurement of Meaning durch Verwendung der stringing-out-Technik Beurteiler- und Konzept-Varianz noch gleichermaßen in die Datenanalyse einbeziehen

(„since our long run purpose was to set up a semantic measuring instrument which would be applicable to people and concepts in general, we wished to obtain that matrix of intercorrelations among scales which would be most representative or typical“, Osgood et al. 1957, 35),

gehen die späteren Arbeiten (insbes. Osgood et al. 1975) davon aus, daß die Beurteilervarianz Vernachlässigbar ist. Demgegenüber erwies sich die Replizierbarkeit der EPA-Struktur nicht nur von einer bevorzugten Berücksichtigung der Konzeptvarianz in den Skalenterkorrelationen abhängig, sondern vor allem auch von der Organisation der Konzeptvarianz durch Art und Anzahl der zur Beurteilung vorgelegten Konzepte. In einer großen Zahl von Untersuchungen werden von der EPA-Struktur abweichende Lösungen berichtet, wenn weniger umfangreiche oder spezifische Konzeptklassen analysiert werden. Derartige Abweichungen lassen sich selbst dann beobachten, wenn die EPA-repräsentativen Beurteilungsskalen verwendet werden. Wenn, wie man vermuten kann, die Skalenterkorrelationen angesichts der meist sehr begrenzten Umfänge der Konzeptstichproben in vielen Fällen nach der stringing-out Methode berechnet wurden, könnte der Einwand von Miron (1972) zutreffend sein, daß in diesen Fällen die Beurteilervarianz eine geringe Kon-

zeptvarianz 'überschwemmt' habe. Allerdings ist dieser Einwand problematisch, weil er die Abweichung durch eine Ursache erklärt, die die Annahme substantiell bedeutungsloser Beurteilervarianz gefährdet.

Gravierender als diese Vermutung erscheint die Kritik, die Carroll schon 1959 gegen die Ergebnisse von Osgood et al. (1957) vorgebracht hat: Im explorativen Zusammenhang muß die verwendete Stichprobe von Konzepten gewährleisten, daß die Skalen entsprechend der gesuchten 'wahren' dimensional Struktur kovariieren. Der gesamte Raum muß entsprechend durch eine angemessene Stichprobe von Konzepten repräsentiert werden; zur Begründung von m unabhängigen Dimensionen fordert Carroll, daß mindestens $2m + 2^m$ Konzeptbeurteilungen zugrundegelegt werden. Dieser Forderung entsprechen die Analysen des Measurement of Meaning offenkundig nicht hinreichend, da die Verwendung von 20 Konzepten danach allenfalls zur Interpretation von vier unabhängigen Faktoren führen kann. Aus diesem Grunde haben Osgood et al. (1975) die Zahl der Konzepte auf 100 erhöht.

Soweit die gesuchte Urteils-Struktur nicht schon als eine inhärente Ordnung der Skalenkovariation angesehen und entsprechend - konzeptfrei - begründet wird (vgl. Analyse 2 des Measurement of Meaning), kann sie - anhand von Konzeptbeurteilungen - nur dann identifiziert und repliziert werden, wenn eine nach Umfang und Art hinreichende Stichprobe von Konzepten zugrunde gelegt wird. Lösungen für einzelne Konzepte oder Konzeptklassen lassen Rückschlüsse auf eine generelle Struktur des Urteilsverhaltens nicht zu.

In welcher Weise beschreibt aber eine konzeptunabhängige, generelle EPA-Struktur die Bedeutung der einzelnen Konzepte, wenn deren Bedeutung beim gleichen methodischen Zugriff durch andere als die EPA-Dimensionen bestimmt wird?

Methodisch entspricht diese Differenz jener zwischen gemeinsamer und spezifischer Kovarianz der Skalen für die Konzepte. EPA repräsentiert die allen (verbalen) Zeichen gemeinsamen Bedeutungsaspekte. Die in sie nicht eingehende, ausgesonderte spezifische Urteilsvarianz führt bei der Analyse einzelner Konzepte/Konzeptklassen zu entsprechend spezifischen Lösungen. Der empirische Gehalt der EPA-Konzeption bemißt sich nicht zuletzt danach, in welchem Ausmaß die auf EPA entfallende gemeinsame Varianz im Verhältnis zur konzeptspezifischen an der Varianz des Urteilsverhaltens beteiligt ist.

In der Tradition der SD-Forschung ist allerdings die Frage, wie gemeinsame und spezifische Bedeutung begrifflich zu fassen sind, von wesentlich größerem Interesse gewesen. Die wichtigste, keineswegs völlig entsprechende Unterscheidung ist die zwischen konnotativer und denotativer Bedeutung.

2.1.2.3 EXKURS: Affektive (konnotative) und denotative Bedeutung

Osgood selbst hat immer wieder nachdrücklich betont, daß EPA Bedeutung nicht in erschöpfender Weise repräsentiert. Ihre Eigenart wird begrifflich als konnotative (emotive, metaphorische) von denotativer (designativer, referentieller) Bedeutung abgegrenzt (Osgood et al. 1957, 321). Die Verwendung dieses in Sprachphilosophie, Linguistik und Psychologie vielfältig variierten Begriffsdualismus durch Osgood hat insbes. von linguistischer Seite heftige Kritik erfahren und zwar sowohl im Hinblick auf ihre begriffliche Klarheit wie auch ihren empirischen Gehalt (vgl. Nordenstreng 1969).

Osgood hatte zunächst erwartet, die zeichenspezifischen Vermittlungsprozesse r_m als Dimensionen sensorischer Art, wie z.B. (visuell) Helligkeit, Farbton und -Sättigungsfaktoren, (auditiv) Lautstärke und Tonhöhe, sowie olfaktorische Faktoren identifizieren zu können. Er bekennt 1964, daß diese Erwartung seiner eigenen mediationstheoretischen Konzeption nicht entsprach, in welcher Bedeutung durch response-spezifische Vermittlungsprozesse definiert wird. Der reaktive Charakter der wiederholt beobachteten EPA-Dimensionen ließ Osgood nicht nur die bessere Übereinstimmung mit seiner Theorie erkennen, sondern auch die affektive Natur dieser Bedeutungsprozesse deutlicher hervortreten (vgl. dazu auch Ertel 1964). Als Reaktion auf einen Beitrag von Kuusinen (1969) tendiert er 1969 dazu, die durch philosophische Sprachverwendung belastete Unterscheidung von Konnotation und Denotation zu verwerfen und von ‚affektiven‘ und ‚anderen‘ Bedeutungskomponenten zu sprechen.

Diese neue Kennzeichnung von Bedeutungsarten erfolgt auf empirischer Grundlage: Die beobachteten EPA-Faktoren weisen bemerkenswerte Ähnlichkeit mit den Wundtschen Hauptrichtungen der Gefühle (Lust/Unlust, Erregung/Beruhigung, Lösung/Spannung) und den - allerdings davon nicht unabhängig formulierten - Dimensionen des mimischen Ausdrucks nach Schlosberg (1954) auf. Über diese begriffliche Analogie hinaus legt Osgood (z.B. 1969, 1971) seiner Interpretation der EPA-Bedeutung Beobachtungen der Dominanz und der Universalität der EPA-Dimensionen zugrunde. Insbesondere aufgrund ihres universellen, kulturunabhängigen Auftretens wird die EPA-Struktur mit neurophysiologischen Spekulationen als angeboren interpretiert; der Dominanz entspreche die fundamentale, ursprüngliche Bedeutung von affektiven Reaktionen für das menschliche Verhalten und zwar sowohl im phylo- wie im ontogenetischen Sinne. In funktionalistischer Betrachtungsweise werden die durch EPA charakterisierten Bedeutungsdimensionen als zentrale Modi der Umweltbewältigung aufgefaßt.

Die operationale Trennung von EPA- und anderen Faktoren, sowie der Hinweis auf eine bemerkenswerte Ähnlichkeit der ersteren mit Emotionsdimen-

sionen, kann eine begriffliche Unterscheidung ‚affektiv-konnotativer‘ gegenüber ‚denotativ-anderen‘ Bedeutungsarten nicht ersetzen. Osgood et al. (1975, 393) akzeptieren die Kritik von Nordenstreng (1969; vgl. auch Miron 1969), insoweit er eine klare, nicht-operationale Definition für das verlangt, was mit einem SD erfaßt wird.

Einen älteren Vorschlag Osgoods (1964) aufgreifend, werden ‚konnotativ‘ solche Urteile genannt, in denen die Bedeutung von Zeichen durch Bezug auf ihre (übergeordnete) Klassenzugehörigkeit spezifiziert wird: ein BABY ist klein, und zwar aufgrund des Vergleichs mit Menschen im allgemeinen; als ‚denotativ‘ werden solche Urteile bezeichnet, die relativ zu einem Standard auf der gleichen hierarchischen Ebene abgegeben werden: NINA ist klein, d.h. im Vergleich zu anderen Babies und nicht als Mensch.

Da beim semantischen Differenzieren, wie die Autoren vermuten, Urteile über ein Konzept relativ zu einem übergeordneten Konzept abgegeben werden (s.o.), würden denotative Urteilsweisen zurückgedrängt. Leider begründen die Autoren die angemessene Verwendung dieser linguistischen Unterscheidung im Hinblick auf das SD-Urteilsverhalten nicht. An anderer Stelle (Osgood et al. 1975, 400) wird das Hervortreten der affektiv-konnotativen EPA-Struktur weiter dadurch erklärt, daß durch die Beurteilung der verschiedensten Konzepte mit den gleichen Beurteilungsmerkmalen eine metaphorische Sprachverwendung begünstigt wird (vgl. Osgood 1969, 1971).

Angesichts der begrifflichen Unschärfe der SD-relevanten Bedeutungsarten beschränken sich operationale Bestimmungen affektiver und denotativer Bedeutung bei homogenen Konzeptklassen auf die Trennung von EPA-spezifischen Bedeutungsanteilen als ‚Affekt‘ von der übrigen, interpretierbaren Skalen-Kovarianz als ‚Denotation‘. Kuusinen (1969) identifiziert affektive und denotative Strukturen von Persönlichkeitsbeurteilungen, indem er aus den Interkorrelationen der Skalenmittelwerte die Korrelationen mit den EPA-Markierskalen herauspartialisiert und sowohl die ursprüngliche, wie die Partialmatrix faktorisiert. Das als denotativ bezeichnete Restsystem ist aber nach diesem Verfahren nicht ein von den EPA-Affekt-Dimensionen gesäubertes System, sondern es ist die gesamte Kovarianz der einzelnen EPA-Markierskalen eliminiert. Tzeng (1975, 1977; Tzeng & May 1975) hat deshalb vorgeschlagen, die Trennung durch Konstruktion von orthogonalen Subräumen für Affekt und Denotation im Bedeutungsraum vorzunehmen: Transformations-Matrizen der Rotation von Markierskalen zur (reinen) EPA-Lösung und zur Einfachstrukturui werden dabei auch zur Rotation von Ladungsmustern für die übrigen Skalen verwendet; EPA- und ‚sonstige‘ Bedeutungsanteile der einzelnen Skalen sind in den jeweiligen Subräumen feststellbar. Im Verhältnis zur Elaboriertheit des Verfahrens tritt die bereits konstatierte begriffliche Vagheit besonders deutlich hervor: als Kriterium der Unterscheidung von affektiver und

denotativer Bedeutung fungiert Ähnlichkeit bzw. Nicht-Ähnlichkeit mit der EPA-Struktur.

Die Bemühungen um eine Präzisierung der SD-Bedeutungsprozesse haben bislang einschlägige Beiträge der ‚imagery‘-Forschung kaum berücksichtigt. Paivio (1969; 1971), ihr bedeutendster Vertreter, faßt ‚imagery‘ als ein non-verbales assoziatives Vermittlungssystem zwischen verbalen Reizen und Responses auf, das neben und unabhängig von einem (teilweise mit ihm verbundenen) verbalen, assoziativen System existiert. Beide Kodierungs-Systeme leisten die kognitive Repräsentation unserer Welt. Die Verfügbarkeit des imaginalen Kodierungs-Systems hängt nach Paivio entscheidend davon ab, in welchem Ausmaß ein verbaler Reiz konkret oder abstrakt ist. d.h. sich auf sinnlich erfahrbare Merkmale von Sachverhalten bezieht:

„The hypothesis is that concrete terms such as „house“ derive their meaning through association with concrete objects and events as well as through association with other words, and thereby acquire the capacity to evoke both nonverbal images and verbal processes as associative (meaning) reactions, which could function as alternative coding systems affecting mediation and memory. Abstract terms such as „truth“, on the other hand, derive their meaning largely through intraverbal experiences and more effectively arouse verbal associative than imaginal processes“ (Paivio 1969, 248).

Denotative Bedeutung bezeichnet die Verknüpfung von Vorstellungsbild (Image) und sprachlichem Zeichen. (Mit Bezug auf Staats, 1968, wird denotative Bedeutung als der auf ein sprachliches Zeichen konditionierbare Teil der sensorischen ‚Reaktion‘ auf das (bezeichnete) Objekt aufgefaßt.)

Insofern besteht zwischen der Dimension abstrakt/konkret und der denotativen Bedeutung von Zeichen ein enger Zusammenhang: abstrakte Wörter, d.h. solche ohne bezeichnete Objekte im Unterschied zu konkreten Wörtern rufen keine Vorstellungsbilder hervor und weisen mithin keine denotative Bedeutung auf.

Godfrey & Natalicio (1970) haben im Anschluß an Paivio die Abstraktheit/Konkretheits-Dimension, deren Rolle von Osgood et al. (1975) mehrfach gering veranschlagt wird (z.B. p. 401, vgl. allerdings p. 187), auf die Evaluationsdimension bezogen und ihren Beitrag durch den Titel ‚Evaluation on SD equals abstraction plus error‘ gekennzeichnet. Diesem Befund liegen z.T. sehr hohe Korrelationskoeffizienten für den Zusammenhang von Urteilsvarianz auf E-Skalen relativ zu NonE-Skalen mit Abstraktheits/Konkretheits-Rangordnungen zugrunde: Bei abstrakten Konzepten tritt mehr E gegenüber NonE als bei konkreten Konzepten auf und vice versa. Lohr (1976) hat gezeigt, daß imaginal-denotative Bedeutung (sensorischer Art) und evaluative Bedeutung (emotional sensu Osgood) als distinkte Vermittlungsprozesse voneinander unabhängig und konkurrierend konditionierbar sind.

KostiE & Das (1971) haben versucht, die Art der mit dem SD erfaßten Bedeutung begrifflich durch Ausschluß von nicht erfaßten Bedeutungsaspekten zu präzisieren:

- Die durch die EPA-Faktoren definierte Bedeutung ist nicht-lexikalischer Art, obwohl die Differentialurteile auf lexikalischen Bedeutungen basieren. Osgood selbst hat wiederholt Beispiele für EPA-bedeutungsgleiche aber lexikalisch verschieden bedeutsame Konzepte gegeben.
- Die mit dem SD erfaßte Bedeutung ist begrenzt auf verbal abstrahierbare Bedeutungsaspekte und zwar solche, die vielen Konzepten gemeinsam sind. Idiosynkratische, spezielle und einzigartige Bedeutungsaspekte werden nicht berücksichtigt.
- Die Definition von Bedeutung durch EPA ist eine allgemeine, ohne Einschränkung im Hinblick auf die individuelle oder spezifische Situation. Im sprachlichen Kommunikationsverhalten wird die Bedeutung von Konzepten durch Adjektive spezifiziert; durch Eingrenzung oder Reduktion vom allgemeinen auf das weniger allgemeine erfüllen Adjektive ihre kommunikative Funktion. EPA leistet eine solche Bedeutungsspezifizierung nicht.

2.1.3 Variationen des Modus der Dimensionsanalyse

Bereits Osgood et al. (1957, 31f) haben die Frage gestellt, ob die dominante EPA-Lösung außer von der Art des samplings in den drei Varianzquellen - Personen, Skalen, Konzepte - auch von der Faktorisierungsmethode abhängig ist. Sie vergleichen die Ergebnisse von Centroid- und Rohwertanalysen und konstatieren befriedigende Übereinstimmung (p. 42ff). Orlik (1965; 1967) bemängelt, daß in Rohwertanalysen dieser Art artifizielle Varianz zwischen den Skalen eingeht und in der Regel einen zusätzlichen, Pseudo-Faktor konstituiert. Für mittelwertszentrierte Rohwert-Produktsummen (Kovarianzen) wird die Möglichkeit erwartungstreuer Abbildung psychologischer Merkmalsräume belegt (vgl. dazu auch Revenstorff, 1973a).

Die Ergebnisse des kulturvergleichenden Forschungsprojekts (Osgood et al. 1975), mit dem die Generalität/Universalität der EPA-Struktur begründet wird, basieren auf Hauptkomponenten-Analysen, die zu einer Lösung führen, die der nach der Centroid-Methode entspricht (Ertel 1965 b; vgl. Harman 1970, 174). Die EPA-Struktur erweist sich als relativ stabil, wenn über die Urteiler gemittelte SD-Ratings faktorisiert werden (vgl. Heise 1969, 415) - ‚repräsentative‘ Stichproben von Konzepten und Skalen vorausgesetzt. Auf die Berücksichtigung interindividueller Differenzen wird noch besonders eingegangen.

Die Validität der EPA-Struktur des semantischen Raumes kann nicht überzeugend durch Variationen der faktorenanalytischen Technik belegt werden, wenn das Urteilsverhalten in der Form von SD-Ratings konstant gehalten wird. Osgood et al. (1957, 143ff) berichten über erste Versuche, die EPA-

Faktoren als die zentralen Dimensionen des bedeutungsspezifischen Urteilsverhaltens über Ähnlichkeitsurteile für Konzepte zu validieren. Anderson (1970) hat diesen Gedanken aufgegriffen und zwölf Adjektive (jeweils zwei für EPA charakteristische Paare) im Paarvergleich auf Ähnlichkeit beurteilen lassen. Aufgrund der stress-Werte der MDS nach Kruskal erwies sich die Dreidimensionalität auch des Ähnlichkeitsraumes als begründet, bei der allerdings - wie auch gelegentlich für SD-Faktoren berichtet wird - A und P zusammenfallen. Diesen, von Osgood als ‚Dynamism‘ bezeichneten komplexen Faktor fand auch Arnold (1971) aufgrund von Kruskal-MDS-Analysen, wobei jeweils vier Dimensionen (nicht-euklidischer Metrik) der Unähnlichkeit für eine Nomen-, eine Adjektiv- und eine Verbliste angemessen erschienen. Diese Unähnlichkeits-Dimensionen wiesen zwar substantielle Korrelationen mit den in Hinblick auf E, P und A auch unidimensional skalierten Listen auf, ohne daß jedoch eine dimensionale Korrespondenz auffindbar wurde. Im Unterschied zu den Ergebnissen von Anderson (1970) basieren die Unähnlichkeitsräume bei Arnold (1971) allerdings auf Wortlisten, deren Geeignetheit zur Reproduktion von EPA zweifelhaft ist. Aber auch angesichts weiterer vorliegender Befunde zur Korrespondenz von SD und MDS-Lösungen für Ähnlichkeits- oder Präferenzräume (Nordenstreng 1968; Green et al. 1969; Magnusson & Ekman 1970; Everett 1973; Shikiar et al. 1974; Gärling 1976) kann eine Entsprechung - zumal im Hinblick auf EPA - nicht hinreichend klar festgestellt werden.

2.1.4 Transkulturelle *Stabilität*

Nachdem sich die EPA-Struktur gegenüber Variationen der Skalen-, Konzept- und Personen-Stichproben - soweit sie hinreichend repräsentativ waren - resistent erwiesen hatte und bereits einige Befunde vorlagen, denen zufolge sie auch über Sprachen/Kulturen Geltung zu haben schien (Kumata & Schramm 1956; Kumata 1957; Triandis & Osgood 1958; Suci 1960) initiierte Osgood 1959 ein gigantisches Forschungsprojekt, bei dem die Generalität der EPA-Struktur über Kulturen und Sprachen überprüft werden sollte. Im Unterschied zu den vorliegenden Befunden wurde auf die Verwendung übersetzungsäquivalenter Beurteilungsmerkmale verzichtet, um sprachlich-kulturellen Eigenarten verschiedener semantischer Systeme Rechnung zu tragen.

Insgesamt gingen Daten aus 25 Sprach/Kultur-Gemeinschaften in diese Untersuchung ein; trotz eines Übergewichtes indo-europäischer Sprachen war versucht worden, linguistische und kulturelle Differenzen zu maximieren.

In einem mehrphasigen Auswahlprozeß wurden 100 Substantive als Beurteilungsgegenstände gesammelt, die ein hohes Maß kultureller Allgemeinheit gewährleisten sollten. Diese zunächst amerikanisch-englische Liste wurde über-

setzt und die Nomen wurden an allen Erhebungsorten jeweils 100 Schülern/Studenten zur Charakterisierung durch jeweils ein Adjektiv vorgelegt. In regulärer grammatikalischer und orthografischer (ggf. transkribierter) Form wurden diese Nennungen von Osgood und Mitarbeiter in Illinois ‚blind‘ im Hinblick auf ‚productivity‘ (Auftrittshäufigkeit und Verteilung über die Konzepte, vgl. Abschn. 4.1), sowie ‚Unabhängigkeit‘ von anderen Beurteilungsmerkmalen geordnet. Auf diese Weise wurden für jede Sprach/Kultur-Gemeinschaft die für sie relevanten Beurteilungsmerkmale ausgewählt. Die örtlichen Forschungsgruppen erhielten eine bis zu 70 Adjektive umfassende Liste zurück, die gemäß diesen Kriterien die höchsten Rangplätze einnahmen. Mit Hilfe von 10 unabhängigen Experten wurden zu ihnen Antonyme erhoben und die danach verbleibenden 50 Adjektivpaare (vorläufig) um 10 Paare ergänzt, die - aufgrund der automatischen Selektion ausgesondert - am Erhebungsort für wichtig gehalten wurden. In einem weiteren Schritt wurden sodann von 200 neu-rekrutierten Vpn (ebenfalls männlichen Schülern/Studenten) alle 100 Konzepte mit diesen Merkmalspaaren in der Form von Rating-Skalen beurteilt. Für die Skalen-Interkorrelationen wurden in Illinois sowohl für Sprache/Kultur spezifische, wie auch pankulturelle Faktorenanalysen gerechnet. In eindrucksvoller Weise wird in nahezu allen Analysen, insbes. der pankulturellen, gezeigt, daß EPA nicht nur durchgängig als dominierende Dreier-Struktur auftritt, sondern ihre semantische Ähnlichkeit zwischen Sprachen/Kulturen teilweise bis in übersetzungsäquivalente Beurteilungsmerkmale reicht. Im theoretischen Zusammenhang ist damit die zentrale Hypothese bekräftigt: „... regardless of language or culture, human beings utilize the same qualifying (descriptive) framework in allocating the affective meanings of concepts“ (Osgood et al. 1975, 6). Dieser Sachverhalt ermöglicht nach Meinung der Autoren für alle Sprachen/Kulturen SDS zu entwickeln, die Unterschiede in subjektiven Kulturen vergleichbar machen, wenn ihre Items EPA repräsentieren.

2.1.5 Interindividuelle Unterschiede

Osgood et al. (1975, 364) stellen zu Recht fest, daß die Osgoodsche Bedeutungskonzeption nicht im Widerspruch zur Möglichkeit individueller Unterschiede bei Bedeutungssystemen steht, die auf unterschiedlichen Erfahrungen beim Lernen von Zeichen oder in Unterschieden hinsichtlich Emotionalität, Intelligenz usw. basieren könnten. Die Universalität der EPA-Struktur ist insoweit nicht theoretisch begründet: Die Annahme ihrer Geltung über Personen geht vielmehr auf Beobachtungen zurück, denen zufolge EPA bei Stichproben verschiedener Personen-Kategorien auftritt, so z.B. im Hinblick auf Alter, Geschlecht, Intelligenz, politische Orientierung, Normalität (vgl. die Übersicht bei Osgood et al. 1975, 58ff; Rosenbaum et al. 1971). Die Befunde zur Generalität über sehr unterschiedliche kulturelle Gruppen haben - auch wenn dabei jeweils nur Stichproben von männlichen Schülern/Studenten be-

rücksichtigt wurden - die Vorstellung bekräftigt, daß alle Menschen ein gemeinsames affektives semantisches Bezugssystem teilen.

In einigen Arbeiten (Williams 1972; Denmark et al. 1972) werden zwar Einschränkungen aufgrund sozio-ökonomischer Klassifizierung von Urteilern nahegelegt; es fällt jedoch schwer, die gefundenen Mängel an Übereinstimmung zwischen den sozialen Gruppen zu interpretieren, da sie gering sind oder durch die Auswahl der beurteilten Konzepte begründet sein können. Hinweise auf kulturspezifische Variationen der Dimensionslösungen werden von Tanaka & Osgood (1965) und Tanaka et al. (1963) bei übersetzungsäquivalenten Skalen in der Weise interpretiert, daß bestimmte Beurteilungsskalen entweder im Hinblick auf Konzepte und/oder Personen faktoriell instabil sind (Konzept-Skalen- und Person-Skalen-Interaktionseffekte).

Den Vergleichen inter- und intrakultureller Stichproben von Personen liegt in der Regel der von Osgood und Mitarbeitern bevorzugte Aggregierungsmodus der dreimodalen Ausgangsdaten Summation' d.h. Mittelwertbildung über Personen bei den Konzepten zugrunde.

Wiggins & Fishbein (1969) bezweifeln, daß Befunde wie die hier angeführten geeignet sind, die Frage nach der universellen Geltung der EPA-Struktur hinreichend zu beantworten. Wenn diese Frage nicht nur die Bedeutung habe, ob es ein gemeinsames semantisches Bezugssystem gebe, das die intraindividuelle Struktur einer 'gemittelten pankulturellen Person' widerspiegle, sondern auch, ob dieses Bezugssystem repräsentativ für (intrakulturelle und) individuelle Strukturen innerhalb der Kultur sei, dann müßte eine auf gemittelten Maßen basierende Struktur auch die beste Repräsentation der semantischen Struktur von Individuen innerhalb einer Kultur sein.

Osgood et al. (1975) rechtfertigen die Verwendung von Gruppenmittelwerten nicht nur mit dem Hinweis, daß für ihre Untersuchungen geeignete dreimodale Analyseverfahren nicht verfügbar waren, sondern sie vertreten weiterhin die Meinung, daß diese Datenreduktion im Hinblick auf interindividuelle Varianz angemessen ist. Erkenntnisgrundlagen dafür sind:

- Faktorlösungen für Datenmatrizen nach Summation- und stringing-out-Reduktion zeigen keine bemerkenswerten Unterschiede. (Bei der stringing-out-Prozedur gehen die einzelnen Personen ein, allerdings ist auch hier Person- und Konzept-Varianz konfundiert.)
- Skaleninterkorrelationen über die Konzepte für jede einzelne Person korrelieren untereinander so hoch, daß die Matrix der Korrelationen über die korrespondierenden Zellen (Personen-Matrix) zu Faktorlösungen führt, die als einfaktoriell anzusehen sind (Ware 1958; Tanaka & Osgood 1965).
- Die Eindimensionalität der Person-Varianz erscheint durch die Ergebnisse

neuerdings vorliegender dreimodaler Faktorenanalysen bekräftigt (Levin 1965; Litt 1966; Snyder 1967; Tzeng 1975).

Gegenüber dem zuletzt genannten Punkt verweisen Wiggins & Fishbein (1969) darauf, daß gerade die Ergebnisse der dreimodalen Faktorenanalyse von Levin die Möglichkeit zur Interpretation mehrerer Person-Faktoren aufweisen. Sie selbst führen eine Tucker-Messick-MDS von Ähnlichkeitsurteilen für 15 charakteristische EPA-Skalen durch und erhalten 3 Person-Faktoren. Während sich die Dimensionalität der Skalen beim ersten Person-Faktor („group average space“) bemerkenswert klar durch EPA reproduzieren ließ, ergab die Einlagerung der Personen in den dreidimensionalen Person-Raum kreisförmige Arrangements um den zweiten und dritten Faktor, deren Repräsentation durch 10 kegelförmig angeordnete Vektoren (Idealisierte Personen) angemessen erschien. Die Ähnlichkeitsurteile dieser 10 homogenen Personen-Gruppen ergaben jeweils Faktorenlösungen („viewpoints“), die zwischen 2 und 4 Skalen-Dimensionen (mit jeweils mehr als 90% Anteil an der gesamten Varianz) aufwiesen. Dabei war nicht nur der Beitrag der einzelnen Skalen zur Definition der semantischen Dimensionen unterschiedlich, sondern es konnte auch beobachtet werden, daß die Dimensionszugehörigkeit der Skalen vom Gruppendurchschnitt zu den idealisierten Personen und zwischen den idealisierten Personen keineswegs stabil war. Dabei ist zu berücksichtigen, daß nur Skalen mit „klarer“ EPA-Indikatorfunktion verwendet wurden.

Bei personspezifischen Faktorenlösungen (Meßwiederholungen über Konzepte) war Ertel (1965 b) zwar zum Ergebnis einer personunabhängigen stabilen EPA-Struktur mit invarianter faktorieller Struktur der Skalen gelangt. Allerdings lagen diesen Analysen Erhebungen bei nur vier Personen zugrunde. Crockett & Nidorf (1967) fanden demgegenüber bei zwölf Vpn eine EPA-Lösung nur für fünf Personen möglich, für die übrigen waren zweidimensionale Lösungen angemessen. Keine Vp zeigte die erwartete Gruppierung der EPA-Skalen zu separaten EPA-Faktoren. Lediglich der E-Faktor war bei allen Personen klar identifizierbar. Hinweise auf interindividuelle Differenzen im Hinblick auf die Struktur des semantischen Raumes finden sich auch bei den Ähnlichkeitsdaten Andersons (1970) und Q-Faktorenanalysen für einzelne Urteilkonzepte durch Revenstorff (1973 a).

Nun wird man die Feststellung von interindividuellen Differenzen im Prinzip als trivial zur Kenntnis nehmen können, solange ihr Ausmaß nicht spezifiziert ist. Der varianzanalytisch feststellbare geringe Varianzanteil zu Lasten der Beurteiler relativ zu dem aufgrund der Konzepte und Skalen (Fuchs 1973; Revenstorff 1973 a, Schäfer 1975 a) läßt keine hinreichenden Rückschlüsse auf die Stabilität der EPA-Struktur über Personen zu. Auch die Interaktions-Varianzen explizieren nicht den strukturellen Aspekt interindividueller Differenzen. Die aufgrund dreimodaler Faktorenanalysen im Hinblick auf die Personen beobachteten mehrdimensionalen Lösungen weisen, soweit Eigenwert-

verlauf und/oder Varianzanteile mitgeteilt werden, für den ersten (Gruppendurchschnitts-)Faktor den Löwenanteil erklärter Personenvarianz aus. Sie lassen aber kaum mehr erkennen, als daß die EPA-Struktur nicht die beste Repräsentation aller individuellen semantischen Strukturen darstellt (Tzeng 1975, 1977; Snyder & Wiggins 1970; Muthen et al. 1977). Shikiar et al. (1974) konnten zwar die Befunde von Wiggins & Fishbein (1969) bekräftigen, die Nützlichkeit der Berücksichtigung interindividueller Differenzen erwies sich für die Vorhersage politischer Präferenzen allerdings als gering. Idiosynkratische- und Gruppendurchschnitts-E-Maße korrelierten nicht nur hoch untereinander, für keine von fünf idealisierten Individuen war die Korrelation des idiosynkratischen E-Maßes mit einem Maß der Wahlpräferenz höher als die zwischen Durchschnitts-E-Maß und Wahlpräferenz. Selbst wenn man die Voraussetzung eines perfekten Zusammenhangs zwischen evaluativen und Präferenz-Urteilen akzeptiert, trägt auch dieser Befund nur sehr vorläufig zur Würdigung individueller Variation im Hinblick auf die Bedeutungsstruktur bei: Interindividuelle Differenzen sind nur im Hinblick auf die E-Dimension berücksichtigt, und die E-Maße konnten nur im Rahmen von SD-Skalen, die aufgrund von ‚Durchschnitts-Analysen‘ für EPA charakteristisch sind, ‚idiosynkratisch‘ sein.

Osgood et al. (1975, 346) konstatieren, daß die affektiven semantischen Systeme von Individuen innerhalb von Kulturen keineswegs völlig homogen sind. Darüberhinaus kann angesichts der vorliegenden Befunde als gesichert gelten, daß EPA weder kulturspezifisch, noch ein Artefakt der Durchschnittsbildung über Personen ist.

2.2 Interaktionsvarianz: Konzept-Skalen-Interaktion

Bedenken gegen die Annahme einer generellen EPA-Struktur sind seit den grundlegenden Arbeiten mit Beobachtungen begründet worden, über die bereits Osgood et al. (1957) berichten: ‚the meanings of scales and their relations to other scales vary considerably with the concept being judged‘ (p. 187). Während eine Person-Skalen- und/oder eine Person(gruppen)-Konzept-Skalen-Interaktion seltener berichtet wird (Krieger 1963; Tanaka et al. 1963; Nordenstreng 1970; Snyder & Wiggins 1970) - was Wunder angesichts der üblichen Durchschnittsbildung der Urteilsmaße über Personen -, ist die Konzept-Skalen-Interaktion ein bevorzugtes Thema der Kritik an der SD-Technik. Allerdings sind die Schlußfolgerungen aus der Analyse der Konzept-Skalen-Interaktion höchst unterschiedlich: sie reichen von der Auffassung, es handle sich um ein ‚Scheinproblem‘ oder methodisches Artefakt (z.B. Ertel 1965a; Kahneman 1963) bis zum Vorschlag, das SD als Datenerhebungstechnik von seiner bedeutungstheoretischen Grundlage abzulösen (z.B. Darnell 1970). Aufgrund des dominierenden Interesses an der Frage der dimensionalen Struk-

tur des semantischen Differenzierens basieren die Interpretationen dieser konzeptspezifischen Einflüsse - wie auch schon für Personeneinflüsse (interindividuelle Differenzen) festgestellt - in aller Regel auf den Ergebnissen von Dimensionsanalysen und selten auf einer varianzanalytischen Erkenntnisgrundlage. Das Ausmaß dieser ‚Anomalie‘ ist deshalb kaum quantitativ präzisierbar und auch nicht im Verhältnis zur Zufallsvariation testbar.

Konzept-Skalen-Interaktionseffekte manifestieren sich in einer erheblichen Variation der Korrelationen zwischen Skalen bei verschiedenen Konzepten (vgl. z.B. Presly 1969; Burns 1976). Entsprechend werden für die einzelnen Konzepte und Klassen von Konzepten unterschiedliche Ladungsmuster der Skalen, einschließlich verschiedener Faktorenstrukturen berichtet (z.B. Osgood et al. 1957; Osgood 1962, Tanaka et al. 1963; Tanaka & Osgood 1965; Darnell 1966; Kubiniec & Farr 1971; Bynner & Romney 1972; Heskin et al. 1973; Klemnack & Ballweg 1973, Burns 1976).

Es entspricht unterschiedlichen Fragestellungen der SD-Forschung, wenn das Auftreten von Konzept-Skalen-Interaktionseffekten einerseits in seiner Relevanz für die Geltung einer generellen EPA-Struktur und andererseits in bezug auf seine Konsequenzen für die Entwicklung eines generell verwendbaren Instruments zur Bedeutungsdifferenzierung gesehen wird.

Osgood und Mitarbeiter haben den Sachverhalt, daß Konzept-Skalen-Interaktionen (im Unterschied zu Person-Skalen-Interaktionen) auftreten, nicht negiert, sondern betont und durch konzeptspezifische Bedeutungsverschiebungen von Skalen zu erklären versucht: In Übereinstimmung mit dem Kongruenzprinzip tendierten im menschlichen Urteilsprozeß alle Skalen zu Bedeutungsverschiebungen in Richtung auf Parallelismus mit dem dominanten Bedeutungsattribut des Urteilskonzeptes. Evaluative Skalen seien für derartige Verschiebungen in besonderer Weise anfällig, SD-Skalen generell als Funktion der ‚evaluativeness‘ von Konzepten. Osgood et al. (1957, 188 u. 326f) vertreten deshalb - was häufig übersehen wird - die Auffassung, man werde zur Erfassung der Bedeutung von Konzepten solche Skalen verwenden müssen, die EPA in konzept(klassen)-spezifischer Weise repräsentierten. Osgood et al. (1975, 351) stellen fest, daß Bedeutungsmaße für bestimmte Konzepte zwischen verschiedenen Kulturen aufgrund von Konzept-Skalen-Interaktionen „cum grano salis“ zu interpretieren seien.

Ertel (1965 a) hat das Auftreten von Konzept-Skalen-Interaktionen einer methodologischen Kritik unterzogen. Er geht davon aus, daß konzeptspezifische Korrelationen keine hinreichende Begründung für Schlußfolgerungen auf eine zugrundeliegende allgemeine Dimensionsstruktur liefern können. Vielmehr könne sich die dimensionale Unabhängigkeit der Urteilsfaktoren nur durch gezielte Variation der Urteilsgegenstände nach dem Kriterium ‚dimensionaler Repräsentativität‘ der Konzeptstichprobe erweisen. Bei einer derartigen Stich-

proben-Organisation würden begriffsspezifische Korrelationen ‚intersituativ‘ verschwinden. Die konzeptspezifische Kovariation der Skalen wird danach verursacht durch irrelevante und störende Bedeutungsaspekte von Konzepten und/oder Skalen. Ähnlich äußert sich Revenstorff (1973a), der das Interaktionsproblem für trivial hält, da man durch eine nicht-repräsentative Auswahl von empirischen Objekten in jeder Korrelationsellipse Ausschnitte wählen kann, die die Korrelation beliebig variieren lassen.

Nun wird man zwar den frühen Arbeiten von Osgood und Mitarbeitern vorhalten können, daß die Anzahl der berücksichtigten Konzepte zu gering war und die Konzepte auch nicht dimensional repräsentativ ausgewählt waren (vgl. die Kritik Carrolls 1959). Immerhin haben sie zu sehr ähnlichen Ergebnissen geführt wie die Ertelschen Arbeiten, in denen die Bedeutungsstruktur aufgrund eines Prozesses sukzessiver Approximation der Auffindung dimensional reiner Skalen und Konzepte als eine mit Erregung, Valenz und Potenz bezeichnete Konfiguration resultierte. Die Forderung nach einer breiteren Urteilsbasis, auch i.S. eines ‚Wechsels der Situation‘, haben Osgood et al. (1975) im Cross-Cultural Projekt erfüllt. Konzept-Skalen-Interaktionseffekte lassen sich dennoch in der beschriebenen Weise beobachten. Ertel selbst hat die Annahme ihres Verschwindens bei dimensional-repräsentativer Konzept- und Skalenauswahl empirisch nicht überprüft. Die Forderung nach repräsentativer Konzeptauswahl ist im übrigen für die meisten Anwendungsfälle der SD-Technik nicht vertretbar.

Während der Zugang Ertels eher das Problem der Identifizierung der Struktur des semantischen Differenzierens und der ihr zugeordneten Skalen betrifft als die Lösung des Problems der Konzept-Skalen-Interaktion, sind andere methodisch begründete Argumente vorgetragen worden, die diese Effekte als ein Artefakt erscheinen lassen. Kahneman (1963) hält die psychologische Erklärung, insbesondere die Annahme konzeptspezifischer Bedeutungsverschiebungen der Skalen für unnötig. Er geht davon aus, daß jedes Rating s_{ijk} bei Konzept j durch Person i auf Skala k als Summe dreier Komponenten aufgefaßt werden kann: dem ‚wahren‘ Wert (Mittelwert der Personen-Population) auf Skala k bei Konzept j , der konstanten Abweichung der Person i auf Skala k und einer spezifischen Abweichung von Person i auf Skala k bei Konzept j , die ihrerseits eine konsistent-idiosynkratische und eine Fehlerkomponente aufweisen. Alle diese Komponenten sind nach Kahnemans Befund korreliert, einschließlich der ‚Fehler‘-Abweichungen, und weisen eine ähnliche Struktur auf. Kahneman kann die in seinen Daten (allerdings in nur geringem Ausmaß) beobachteten Konzept-Skalen-Interaktionseffekte im wesentlichen durch die idiosynkratische Komponente der spezifischen Abweichung in Form einer konstanten Überschätzungs/Unterschätzungs-Tendenz des wahren Wertes erklären: unterschiedliche Korrelationen (Vorzeichen) zwischen Skalen bei einzelnen Konzepten können infolge dieser interindividuellen Unterschiede er-

wartet werden, je nachdem, ob die wahren Werte auf der gleichen oder verschiedenen Seite des Skalenneutralpunktes liegen. Unterschiedliche Skalen-Korrelationen (der wahren Werte) bei verschiedenen Konzeptklassen seien demgegenüber nicht auf kognitive Strukturen der Urteiler, sondern auf reale ‚ökologische‘ Beziehungen zwischen den Konzepten zurückzuführen. Auch für diese Art der Konzept-Skalen-Interaktion wird die Annahme von Bedeutungsverschiebungen der Skalen zurückgewiesen.

Heise (1969) betont, daß Beobachtungen von Konzept-Skalen-Interaktionen in der Regel bei Analysen auf der Basis von Personwerten und nicht von Gruppenmitteln berichtet werden. Daß hierbei person-spezifische Varianz die diskutierte Interaktion zu beeinflussen scheint, wird durch die Arbeit von Snyder & Wiggins (1970) nahegelegt: die Autoren weisen aufgrund einer dreimodalen Faktorenanalyse nicht nur auf eine „interaktionale Beziehung zwischen idealisierten Personen, Konzepten und Skalen“ (p. 466) hin, sie charakterisieren die idealisierten Personen durch Extremisierungstendenzen, die allerdings abhängig von Urteilsdimensionen und Skalen erscheinen. Nordenstreng (1969) hebt hervor, daß die Kahnemansche Unterscheidung des kognitiven und des ökologischen Systems immerhin deutlich macht, daß sowohl Personen wie Konzepte zu Konzept-Skalen-Interaktionseffekten beitragen. Die Feststellung einer psychometrischen Beziehung könne ihre konzeptuelle Interpretation aber nicht ersetzen. Er unterscheidet vier Typen (A-D):

„... different correlations between corresponding scales indicate in type **A** the extent to which a set of concepts is related differently to two individuals, in type **B** the extent to which an individual is related differently to two sets of concepts, in type **C** the extent to which a concept is related differently to two sets of individuals, and in type **D** the extent to which a set of individuals is related differently to two concepts“ (p. 13).

Heise (1969) spricht von ‚wahrer‘ Konzept-Skalen-Interaktion in Abgrenzung von methodisch artifizieller, wie sie durch unangemessene Konzeptauswahl, Verwendung irrelevanter Skalen und Polarisierungsfehler bedingt sei. Wahre Interaktionseffekte könnten durch unterschiedliche Relevanz der Skalen für verschiedene Konzepte und durch Bedeutungsverschiebungen in den Skalen im Hinblick auf Klassen von Konzepten zustandekommen.

Das Merkmal ‚süß‘ - ‚sauer‘ mag hoch relevant zur Beurteilung von (bestimmten) Nahrungsmitteln, mäßig relevant zur Beurteilung von Mitmenschen und kaum relevant zur Beurteilung abstrakter Ideen sein. Entsprechend wird bedeutsame Urteilsvarianz auf dieser Skala geringer werden. Insoweit die Kovariation mit anderen Skalen dadurch beeinflußt wird, variieren - konzeptspezifisch - die Ladungsmuster von Faktorenlösungen.

Eng verbunden mit der konzeptspezifisch (und mutmaßlich auch person-spezifisch) differentiellen Relevanz von SD-Skalen scheint die auf Bedeutungsarten bezogene Erklärung von Verschiebungen der Skalen-‚Bedeutungen‘ durch

Osgood (1962) zu sein. Durch die Integration des affektiv-energetischen und des sensorisch-motorischen Diskriminierungssystems mit den ihnen zugeordneten konnotativen bzw. denotativen Bedeutungsreaktionen in gleichen Systemen des Sprachverhaltens trete in den auf die Erfassung von affektiver (konnotativer) Bedeutung gerichteten Skalen ‚denotative Kontamination‘ auf. Einzelne Konzepte schränken die Skalenbedeutungen in selektiver Weise ein: so werde durch das Konzept LAVA für die Skala ‚heiß - kalt‘ deren Denotation hervorgerufen, während deren (activity) Konnotation durch Konzepte wie JAZZ und FESTIVAL betroffen werde.

Welche Folgerungen lassen sich aufgrund der vorliegenden Befunde aus dem Sachverhalt der Konzept-Skalen-Interaktion im Hinblick auf die Frage nach der Generalität der EPA-Struktur einerseits und der generellen Verwendbarkeit eines SD-Instruments andererseits ziehen?

Zunächst operieren derartige konzept- wie auch personspezifischen Einflüsse gegen die Begründung eines generellen/universellen affektiven Bedeutungssystems. Ihre Auswirkungen können aber angesichts der vorliegenden Befunde (vgl. auch die Ergebnisse der ‚konzeptfreien‘ Skalenstruktur-Analysen bei Osgood et al. 1957; Wiggins & Fishbein 1969) die Feststellung einer stabilen EPA-Kernstruktur des semantischen Differenzierens nicht erschüttern; abgesehen von Meßfehleranteilen verweisen sie vielmehr auf spezifische Bedeutungsaspekte, die außerhalb des Bedeutungshorizonts der EPA-Konfiguration für Konzepte (und Personen) charakteristisch sind. Offenkundig verfügt keine der von Osgood und Mitarbeitern im Cross-Cultural-Projekt berücksichtigen Sprachen/Kulturen über „reine“, konzeptunabhängige EPA-Indikatoren. Die Beurteilungsskalen transportieren jeweils Bedeutungsaspekte über die durch EPA definierten hinaus und zwar in interindividuell und zwischen Konzepten varianter Weise.

Es erscheint gerechtfertigt, diese spezifischen Abweichungen bei der Suche nach einer *allgemeinen* Struktur im Kognitionsverhalten als Fehlerkomponente aufzufassen. Sie scheinen jedoch von hinreichend systematischer Art zu sein, um *spezifische* Bedeutungsstrukturen identifizierbar zu machen (vgl. Kuusinen 1969; Tzeng 1975, 1977; Tzeng & May 1975).

Die Frage nach den Konsequenzen der Konzept-Skalen-Interaktion für Konstruktion und Anwendung eines SD außerhalb des Kontextes der Suche nach einer allgemeinen Bedeutungsstruktur ist weniger eindeutig zu beantworten. Einerseits ist der Wert der SD-Technik wesentlich von der Bezugsmöglichkeit auf die allgemeinen Bedeutungsdimensionen EPA abhängig. Andererseits kann kaum bezweifelt werden, daß konzept- und personspezifische Einflüsse die Bedeutungsdimensionierung in unbestimmter Weise verzerren, wenn sog. generelle Skalen verwendet werden; die dimensionale Repräsentativität auch von ‚typischen‘ EPA-Skalen variiert in Abhängigkeit von diesen Einflüssen.

Aus der Generalität der EPA-Struktur folgt allerdings nicht, daß diese Dimensionen für alle Konzepte und Personen(gruppen) durch die gleichen Skalen optimal repräsentiert werden. Im Gegenteil zeigen die Interaktionseffekte, daß dies nicht der Fall ist. Die sprach/kulturspezifischen Lösungen des Cross-Cultural-Projekts zeigen vielmehr, daß EPA (bei Sprach/Kulturgruppen) durchaus auch im Gewande verschiedener Skalensets auftritt. Eine Lösung des Dilemmas könnte also darin bestehen, für eine bestimmte homogene Klasse von Konzepten und eine bestimmte Population von Personen spezifische EPA-Indikatoren zu suchen und zu verwenden.

Die üblicherweise zur Behandlung von Konzept-Skalen-Interaktionseffekten vorgeschlagenen ‚Lösungsmöglichkeiten‘ betreffen die Gewichtung der einzelnen Skalen für die Berechnung von Person-Maßen für die einzelnen Dimensionen (Faktoren-Scores), die - häufig als Bedeutungs- und Einstellungsmaße verwendet - in besonderer Weise von instabilen Faktorstrukturen und Ladungsmustern beeinträchtigt werden. Presly (1969) hält mit Bezug auf Konzept-Skalen-Interaktionseffekte eine Gewichtung von Skalen-Werten nur auf der Grundlage konzeptspezifischer Faktorenladungen für vertretbar. Bynner & Romney (1972) wollen wenigstens die damit (aufgrund unterschiedlicher Faktorenlösungen) preisgegebene Vergleichbarkeit der Faktoren-Scores in einem gemeinsamen Faktoren-System retten und empfehlen Faktorenanalysen sowohl für die einzelnen wie über alle Konzepte. Soweit Faktoren der Analyse über die Konzepte auch in den konzeptspezifischen Analysen auftreten, sollten für diese Faktoren-Scores nach Maßgabe der Gewichte der Analyse über die Konzepte berechnet werden. Levy (1972) hebt hervor, daß die Identifizierung gemeinsamer und konzeptspezifischer Faktoren erhebliche Probleme aufwerfe und schlägt andere Lösungen vor: Faktorisierung der (Konzepte/Skalen) x (Konzepte/Skalen) Korrelationsmatrix (vgl. auch Kubiniec & Farr 1971; Klemmack & Ballweg 1973); oder: Projektion der Faktorenstrukturen der untersuchten Konzeptklasse in einen Bedeutungsraum für Standard-Konzepte, etwa den EPA-Raum; oder: Verwendung einer dreimodalen Faktorenanalyse-Prozedur. Zur Berücksichtigung von Variation in Mittelwerten und Standardabweichungen der Skalen über die Konzepte wird die Analyse von Kreuzprodukt-Rohwerten und Kovarianzen empfohlen.

Datenanalyseverfahren können Konzept-Skalen-Interaktionseffekte deutlich und lokalisierbar machen. Sie können dazu beitragen, der Konzept-Skalen-Interaktion als einer Anomalie Rechnung zu tragen. Sie können aber weder das Problem störender konzept- und personenspezifischer Einflüsse „lösen“, noch ihr Auftreten beeinflussen. Letzteres kann allerdings durch angemessene Auswahl von Skalen für bestimmte Konzepte (bei bestimmten Personen) erreicht werden (vgl. Abschn. 4).

2.3 „Fehlervarianz“

2.3.1 Systematische Urteilsfehler

2.3.1.1 Extremisierung

Bei der Inspektion von SD-Datenlisten sind in der Regel bei einzelnen Personen auftretende Bevorzugungen bestimmter Skalenpositionen, und zwar der Mittel- und der Extrempositionen auffällig. Derartige Urteilstendenzen sind für Rating-Skalen häufig beschrieben worden (z.B. Peabody 1962; Hamilton 1968). Wir beschränken uns hier auf einige Beobachtungen und Erklärungen dieses Sachverhalts bei der Analyse von SD-Daten. Dabei steht die ‚Extremisierungstendenz‘ im Vordergrund. Die Bevorzugung der mittleren Antwortkategorie scheint nach den vorliegenden Befunden (Mitsos 1961; Orlik 1965; Mikula & Schulter 1970; Grimm et al. 1973) von der Auswahl der Skalen für die Konzeptbeurteilungen abzuhängen.

Herrmann (1962) berichtet über zwei voneinander unabhängige Tendenzen der Präferenz von Skalenstufen: „Urteilsnuanciertheit“ als Wahl der Stufen 1 und 7 gegenüber 3 und 5 und „Fraktionierung des Bezugssystems“ als Wahl von 4 gegenüber 2 und 6 bei der siebenstufigen Skala. Diese Faktoren des semantischen Differenzierens werden als genuine Aspekte von Bedeutungsurteilen und nicht als Fehlerkomponenten aufgefaßt. Die Diskussion Herrmanns legt eher eine Abhängigkeit von der Art der beurteilten Konzepte als eine differentiell-psychologische Interpretation nahe. Demgegenüber resümiert Peabody (1962, 73):

„Individual differences in average extremeness show wide generality across different extremeness scores. This generality extends to responses in opposing directions - suggesting that the differences primarily represent response sets, and only to a secondary degree actual differences in intensity.“

Wenn man Gruppenmittel als ‚wahre Werte‘ betrachtet, lassen sich - wie Kahneman (1963) gezeigt hat - die individuellen Abweichungen bei SD-Daten als Über-/Unterschätzungstendenzen interpretieren. Osgood et al. (1957, 226 ff) diskutieren sehr ausführlich Befunde, derartige Urteilstendenzen als persönlichkeitspezifische Stile aufzufassen. Die in neueren Untersuchungen berichteten Beziehungen zu Merkmalen wie Alter und Geschlecht (Maltz 1963; Light et al. 1965; Long et al. 1968; Schludermann & Schludermann 1969; Washington 1975), IQ (Neuringer 1963; Brod et al. 1964; Light et al. 1965; Stricker & Zax 1966; Long et al. 1968), Ambiguitätstoleranz (Mogar 1960; Rydell 1966; Chen 1977), psychischen Störungen (Neuringer 1963; Zax et al. 1964; Marks 1965; Arthur 1966; Priest 1971) sind allerdings uneinheitlich. Washington (1975) stellt fest, daß aufgrund unterschiedlicher methodischer Ansätze die Vergleichbarkeit der Befunde im Hinblick auf verschiedene Kon-

zepte und Skalen kaum gegeben sei. Er empfiehlt, entsprechende Analysen für einzelne Konzepte und dimensionsgleiche Skalen (auch) getrennt durchzuführen. Zwar hat Arthur (1966) bei einer studentischen Vpn-Stichprobe hohe Korrelationen für Häufigkeiten extremer Markierungen über ein Vierwochen-Intervall und über verschiedene Konzepte gefunden. Die Stabilität der Extremisierungstendenz scheint nach dem Befund von Snyder & Wiggins (1970) aber sowohl im Hinblick auf Skalen als auch Konzepte eingeschränkt: die Differenzierung zweier Personentypen (als Ergebnis einer dreimodalen Faktorenanalyse) basierte auf Überbewertungen bei den E-Skalen durch die erste idealisierte Person, während der zweite Personentyp durch ähnliche Tendenzen bei den P-A-Skalen charakterisiert war. Allerdings waren diese allgemeinen Tendenzen von Konzept zu Konzept verschieden. Long et al. (1968) zufolge verwenden Schüler bei Selbst-Ratings für E- und P-Skalen unterschiedlich stark Extremkategorien, und zwar Mädchen stärker bei E und Jungen stärker bei P.

Personen(gruppen) unterscheiden sich im Ausmaß der Verwendung extremer Skalenpositionen. Obwohl diese Tendenz systematisch zu sein scheint, lassen sich ihr keine spezifischen Persondispositionen zuordnen. Sie scheint von Konzept-Skalenauswahl abhängig zu sein. Heise (1969) schlägt vor, ihr (in Experimenten) durch randomisierte Zuordnung der Vpn oder durch statistische Kontrolle Rechnung zu tragen. Johnson & Wall (1969) kritisieren die Anwendung von ‚Ausfilterungs-‘ und anderen statistischen Korrekturtechniken: Wenn die Extremheit eines Urteils zwar nicht eindeutig als Intensitätsmaß aufgefaßt werden kann, andererseits aber nicht unabhängig von der Bedeutung eines beurteilten Konzeptes variiert, wird durch derartige Techniken die Bedeutungseinschätzung eines Konzeptes mitverändert.

2.3.1.2 Soziale Erwünschtheit

Die Kontamination von Urteilsverschiebung und Bedeutungsaspekten wird in einer systematischen Urteilstendenz deutlich, die empirisch nicht immer ohne nähere Analyse von der Extremisierungstendenz zu unterscheiden ist (vgl. die zitierte Untersuchung von Long et al. 1968) und als Interpretationsalternative bei Bedeutungsmaßen die Validität einer SD-Bedeutungsinterpretation in Frage stellen kann: Die Extremheit der Urteile hängt auch davon ab, wie sehr die Konzepte und Skalen sozial normierte Beurteilungssachverhalte und -gesichtspunkte betreffen. Sozial positiv bewertete Wörter weisen höher polarisierte Bedeutungsmaße auf als negative, wobei Polarisierung und Evaluation konfundiert erscheinen (Howe 1965). Nickels & Shaw (1964) stellten fest, daß die Korrelation zwischen E-Faktoren-Scores (als Einstellungsmaßen) und Maßen von Thurstone-Skalen je nach dem Grad der Bedeutsamkeit (salience) der beurteilten Konzepte für Beurteiler variierte. Personen seien sozialen Normie-

rungen des Urteilsverhaltens bei salienten Konzepten stärker ausgesetzt und die wirkten sich beim transparenteren Einstellungsmaß, dem SD, stärker aus.

Krieger (1963) hat die soziale Erwünschtheit des Urteilsmediums ‚Skalen‘ kontrolliert; sie fand für balancierte Skalen geringere Beiträge zur E-Dimension. Ford & Meisels (1965) beziffern das Ausmaß der Korrelation zwischen Maßen sozialer Erwünschtheit von SD-Adjektivpaaren und Ladungen auf der E-Dimension mit .88 bis .92, bei Ladungen auf P und A dagegen mit nur .13 bis .22. Nun belegen derartige Befunde lediglich, daß für die Bedeutungsdifferenzierung auf der E-Dimension (im Unterschied zu P-A) Merkmale konstitutiv sind, die Sachverhalte nicht unwesentlich durch die unterschiedliche Zuordnung (sozial) erwünschter Eigenschaften charakterisieren. Diese empirische Korrespondenz impliziert noch keineswegs die Feststellung systematischer Urteilsverzerrungen, wie Heise anzunehmen scheint. (Revenstorff (1971, 196) hält gar - konsequenterweise - das Vorkommen eines evaluativen Faktors als nachteilig für den Vergleich von Persönlichkeitsbeschreibungen.) Als Tendenzfehler ist die (differentielle) Abhängigkeit des Urteilsverhaltens von der sozialen Erwünschtheit der Konzepte und/oder der Skalen zu belegen.

Meisels & Ford (1969) haben in einer weiteren Arbeit personspezifische Urteilstendenzen i. S. sozialer Erwünschtheit nachzuweisen versucht. Die berichteten Korrelationskoeffizienten zwischen verschiedenen derartigen Tendenzmaßen und Maßen für EPA bestätigen die Unabhängigkeit von P und A und lassen - entgegen der Interpretation der Autoren - allenfalls eine schwache Beziehung zu den E-Maßen erkennen.

Wenngleich ungewöhnliche personspezifische Tendenzen, Urteile gemäß ihrer sozialen Erwünschtheit abzugeben, für die SD-Technik nicht belegt sind, wird man deshalb eine entsprechende Fehlerquelle nicht negieren können (vgl. dazu Joyce & Jackson 1977). Insbesondere bedarf der Zusammenhang von sozialer Erwünschtheit mit der E-Dimension einer näheren Begründung.

Für die Erklärung eines möglichen Tendenzfehlers der beschriebenen Art könnte der Bezug auf die in der Tradition der Einstellungsforschung und der Urteilstheorien untersuchten Polarisierungseffekte nützlich sein. Beurteiler, für die ein Urteilsgegenstand ausgeprägten Wertbezug aufweist, tendieren dazu, die (Un)Günstigkeit von Feststellungen über diesen Sachverhalt polarisierter zu beurteilen (Hovland & Sherif 1952; Zavalloni & Cook 1965; Eiser & Stroebe 1972; Eiser 1971a, 1971 b; Eiser & Mower White 1974). Jones (1969) hat gezeigt, daß die Bedeutung des am meisten akzeptierten Beurteilungsstatements erheblich stärker durch die E-Komponente charakterisiert ist als die Bedeutung des am meisten abgelehnten - ein Hinweis, daß positiver Wertbezug sich wie social desirability und evtl. auch saliency durch Polarisierung auf der E-Dimension auswirkt. Im Unterschied dazu stellte Jones signifikant höhere Ratings bei den abgelehnten Statements auf der P-Dimension fest.

Auch die Theorien zur Wahrnehmungs- und sozialen Akzentuierung (vgl. Irle 1975; Lilli 1975), so die Reizklassifikationstheorie Tajfels (1959, 1975), betonen die Polarisierung von Urteilen in Abhängigkeit von einem Wertbezug.

2.3.2 Zufallsfehler - Reliabilität von SD-Urteilen

Osgood et al. (1957, 126ff) haben mit großer Sorgfalt die Reproduzierbarkeit der wichtigsten Bedeutungsmaße aufgrund von Meßwiederholungen untersucht. (Wegen der z.T. geringen Streuungen der Urteilsmaße über die Personen halten sie die Korrelationstechnik *zur* Reliabilitätseinschätzung für ungeeignet.) Sie unterscheiden die Aspekte der Reproduzierbarkeit auf den Ebenen der Skaleneinstufungen und der Faktoren-Scores, sowie der von letzteren abhängigen Konzept-Bedeutungsmaße im semantischen Raum. Die berichteten 5%-Signifikanzgrenzen für Differenzen zwischen je 2 Erhebungen liegen für die einzelnen Personen und Skalen (7 Stufen) bei mehr als 2 Skaleneinheiten, für Faktoren-Scores einzelner Personen zwischen 1 und 1,5 Skaleneinheiten (je nach Faktor; E am stabilsten) und für gruppenspezifische Faktoren-Scores (Gruppenmittel) bei etwa 0,5 Skaleneinheiten. Die mittleren absoluten Abweichungen zwischen 2 Erhebungen bei den verschiedenen Skalen liegen im Durchschnitt bei etwa $3/4$ Skaleneinheiten (vgl. Osgood et al. 1957; Piaggio 1968).

Da der Spielraum für Diskrepanzen zwischen den Meßwiederholungen von der Extremität der ersten Messung abhängig ist, bezieht Norman (1959) die beobachteten Diskrepanzen auf die maximal möglichen. Die auch schon von Osgood et al. berichteten geringen Konsistenzen der einzelnen Ratings lassen sich nach Norman durch Faktoren-Scores leicht verbessern, wenn mindestens drei Skalen zugrunde gelegt werden. (Dieser Befund basiert allerdings auf Daten, in denen die Skalenzahl nicht unabhängig von der Art der Bedeutungsdimensionen variiert wurde. Im übrigen läßt sich der Meßfehleranteil natürlich durch ‚Test‘verlängerung reduzieren (Piaggio 1969). Bei dimensional repräsentativ ausgewählten Skalen sind allerdings in der Regel nur wenige geeignete Indikatoren verfügbar; auch dürfte der Gewinn gering sein.) Während die D(istanz)-Maße für einzelne Skalen und Personen zwischen Konzepten extrem geringe Stabilität aufwiesen, korrelierten D-Maße, gemittelt über die Personen, sowie gemittelte Skalenmarkierungen zwischen den beiden Messungen hoch. Die Stabilität von Ratings ist nach Norman sowohl bei verschiedenen Konzepten wie bei verschiedenen Personen unterschiedlich.

DiVesta & Dick (1966) haben eine umfangreiche Untersuchung zur Reliabilität von SD-Daten bei Schulkindern durchgeführt, mit zwei Messungen in unmittelbarem zeitlichem Zusammenhang bzw. im Abstand von etwa vier Wochen. Unter der Bedingung des 4-Wochen-Intervalls betrugen die durchschnittlichen Korrelationskoeffizienten für die einzelnen Skalen über die Schulklassen .27

bis .56 und für einzelnen Klassen aufsteigend von .33 (in der 2. Klasse) bis .55 (7. Klasse). Die Höhe der Koeffizienten stieg bei Addition von jeweils zwei dimensionsgleichen Skalen-Werten zu Faktoren-Scores, nunmehr zwischen 35 und .86 variierend, wobei ab der 4. Klasse eine deutliche Steigerung zu beobachten war. Korrelationen für Maße der Bedeutungs sättigung im semantischen Raum lagen etwa im Bereich von .50 bis .70. Für diese insgesamt unbefriedigenden Ergebnisse werden von den Autoren Bedeutungsänderungen der Konzepte zwischen den Meßzeitpunkten geltend gemacht: Die Koeffizienten waren für Erhebungen mit einem Zeitintervall von einem bis zwei Tagen erheblich höher. Für die untersuchten Klassenstufen 3, 5 und 7 betrugen die durchschnittlichen Korrelationen über die Skalen .56, .56 bzw. .67, für die einzelnen Skalen .42 bis .77. Faktoren-Scores für Personen korrelierten bei den EPA-Faktoren zwischen .62 und .84 (am höchsten für E) und für Konzepte gemittelt über Personen zwischen .73 und .94. Entsprechend erhöht sind auch die Korrelationen für die Bedeutungs-Distanzmaße. Die Autoren bewerten diese Ergebnisse als Beleg für akzeptable Stabilität von SD-Daten, eine Würdigung, die allenfalls angesichts des Alters der untersuchten Personen akzeptabel erscheinen mag.

Eher befriedigende (Alpha)Koeffizienten werden von Oles (1973) berichtet: sie liegen für jeweils sechs E-Skalen über neun Konzepte bei Dritt- bis Fünftkläßlern bei .86 bis .92 und summiert über die Skalen innerhalb der Konzepte bei .54 bis .72. Als Stabilität über sieben Monate sind - angesichts der untersuchten Altersgruppen immerhin noch - Koeffizienten von 35 bis .44 für Konzeptsummen bei bedeutungsstabilen Konzepten ausgewiesen.

Miron (1961) betont zu Recht, daß Einschätzungen von Eigenschaften eines SD dem Umstand Rechnung tragen müssen, daß ein SD nicht als ein spezifischer Test, sondern als eine Technik anzusehen sei. Zu den variablen Merkmalen dieser Technik gehören auch die Instruktionsbedingungen. Er variiert systematisch die Faktoren Schnelligkeit der Bearbeitung und Rekapitulation der Markierungen des ersten Durchgangs (jeweils in zwei Stufen) im unmittelbar darauffolgenden Retest. Faktoren-Scores über drei bis fünf Skalen für EPA-Faktoren, gemittelt über die Vpn, korrelieren über 20 Konzepte unter allen Bedingungen mit mindestens .97. Eine Varianzanalyse der Test-Retestabweichungen für die gemittelten Konzept-Scores weist einen signifikanten Haupteffekt zu Lasten der Rückruf-Bedingung: die Vpn konnten, wenn sie entsprechend instruiert wurden, Markierungen erinnern. Dieser Befund läßt die Bevorzugung der unmittelbar erfolgten Retest-Messung bei der Interpretation der Reliabilität von SD-Messungen durch DiVesta & Dick (1966) als problematisch erscheinen und belegt die Berechtigung der von Gulliksen (1958) bereits formulierten Forderung, Parallelversionen zu verwenden; in dieser Hinsicht ist die Arbeit von Coyne & Holzman (1966) eine Ausnahme geblieben.

Abweichungen der Skalen von klarer Bipolarität (fragwürdige Gegensätzlichkeit der polaren Adjektive) und Konzept-Skalen-Kombinationen, die das Auftreten von Interaktions-Effekten begünstigen, scheinen das SD-Urteilsverhalten kaum zu beeinflussen. Vidali (1976) fand unter diesen Bedingungen keine nennenswerten Unterschiede bei Reliabilitätsmaßen für einzelne Rater (um .50) und für Gruppen von Urteilern (um .97). Maruyama (1971) zeigte, daß die mittlere Antwortkategorie der Skalen Reliabilitäts- und Stabilitätsmaße beeinflussen kann: Die von ihm berechneten Koeffizienten waren höher, wenn o-stufige Skalen (ohne Mittelpunkt) verwendet wurden, im Vergleich zu 7-stufigen Skalen.

Auf Reliabilität im Sinne hoher Stabilität und Reproduzierbarkeit der Faktorladungen für die charakteristischen EPA-Skalen, und zwar für Korrelationen individueller wie über Personen gemittelter Ratings bei 25 Konzepten und einer repräsentativen amerikanischen Stichprobe, verweist Tzeng (1975). Daß hohe Skalen-Homogenität, als Reliabilitätsmaß ausgedrückt durch einen Generalisierbarkeits-Koeffizienten (vgl. Gleser et al. 1965), erreichbar ist, belegen die konzept- und personspezifisch konstruierten SDs von Fuchs (1973) und Schäfer (1975a, 10).

Es ist Fuchs (1975, 84f) zuzustimmen, der die Ergebnisse der einschlägigen Arbeiten folgendermaßen zusammenfaßt.

- (1) „Geht man von den einzelnen Beurteilern aus, ist die Reliabilität unter allen Aspekten - obwohl deutlich besser als nach dem Zufall zu erwarten wäre - nicht zufriedenstellend. Geht man dagegen von Gruppenmitteln aus, erhält man - verglichen mit anderen subjektiven Tests - sehr zufriedenstellende Reliabilitätswerte.
- (2) Es gibt deutliche Unterschiede zwischen Beurteilern (z.B. urteilen ältere Kinder konsistenter als jüngere), Skalen (z.B. werden Skalen der Bewertungsdimension konsistenter verwendet als andere) und Konzepten (z.B. werden ‚objektive Konzepte‘ konsistenter eingeschätzt als ‚subjektive‘, ‚nicht-neutrale‘ konsistenter als ‚neutrale‘).
- (3) Dimensionswerte, d.h. über alle auf einen bestimmten Faktor hoch und möglichst rein ladenden Skalen - evtl. unter Berücksichtigung der Ladungshöhe - gebildete Konzeptmeßwerte, sind zuverlässiger als Item-Einstufungen . . .
- (4) Die Reliabilität nimmt offensichtlich ab mit dem zeitlichen Abstand zwischen Test und Retest.“

3. *Metrische Eigenschaften von SD-Skalen: 'Statik' des semantischen Raumes*

Es ist bereits erwähnt worden, daß die SD-Technik meßtheoretisch zunächst keinen anderen Status beanspruchen kann als andere Rating-Verfahren: Es handelt sich um Messungen ‚per fiat‘, numerische Indizierung ohne Lösung des Repräsentationsproblems. Im Hinblick auf die Vertretbarkeit von Transformationen im Prozeß der Analyse von SD-Daten (einen Überblick über die gängigen SD-spezifischen Techniken geben Diehl & Schäfer 1975) sind bereits von Osgood et al. (1957) einige metrische Eigenschaften der Skalen diskutiert worden, die für den Architekten eines semantischen Raumes als Probleme der Statik gelten können. Zu den grundlegenden metrischen Annahmen, die als Merkmale der Skalen Eigenschaften des semantischen Raumes betreffen, gehören Bipolarität, Intervallgleichheit und Nullpunktlage gemäß der numerischen Kodierung.

3.1 Bipolarität

Mit Bezug auf die Charakterisierung der bedeutungsspezifischen Vermittlungsreaktionen als reziprok-antagonistisch ist für die Bedeutungsdimensionen und die sie konstituierenden Skalen zu fordern, daß ihre Pole Gegensätze auf eindimensionalen Kontinua repräsentieren.

Kjeldergaard & Higa (1962) konnten zeigen, daß das Wiedererkennen von Wörtern durch den Grad ihrer Polarisierung im semantischen Raum begünstigt wird. Aufgrund von Analysen der Enkodierung und Speicherung von Wörtern im Kurzzeitgedächtnis sowie des Reproduzierens lassen sich die Pole der EPA-Dimensionen als verschiedene Klassen der Kodierung von Wörtern auffassen (Markel et al. 1966; Wickens & Clark 1968; Wickens 1970; Kroes & Libby 1971). Turvey et al. (1969) und Turvey & Fertig (1970) konnten zeigen, daß die Unähnlichkeit von Wörtern zwischen und die Ähnlichkeit in diesen Klassen auf die Polarität der EPA-Dimensionen zurückgeführt werden kann. Nach Befunden von Haygood (1966) und Taylor & Haygood (1968) wurden semantische Konzepte um so schneller gemäß den EPA-Dimensionen kategorisiert, je stärker die Kategorien polarisiert waren.

Derartige Befunde stimmen zwar mit der Annahme überein, daß die beim semantischen Differenzieren verwendeten Dimensionen als bipolar zu konzipieren sind. Sie erübrigen aber nicht die Prüfung, ob, in welchem Ausmaß und unter welchen Bedingungen Bipolarität als ein Merkmal von SD-Skalen und -Dimensionen gilt.

Aufgrund von Analysen der Assoziationen für die häufigsten englischen und die Standard-SD-Adjektive kommt Deese (1964) zu dem Ergebnis, daß das

Schema polarer Gegensätze -wenn auch nicht für alle Adjektive - durchaus im Sprachverhalten begründet ist. Die Ergebnisse weisen die SD-Polaritäten überwiegend als linguistische Kontraste aus. Carter et al. (1969), die ihre Vpn auch aufforderten, das Antonym zur nur einseitig markierten Skala einzusetzen, fanden für die 15 am höchsten ladenden SD-Skalen in der Mehrzahl die von Osgood et al. (1957) verwendeten wieder, in anderen Fällen aber bemerkenswerte Abweichungen. Ross & Levy (1960) bezweifeln, daß Adjektive polarer Anordnung im Hinblick auf ihre semantische Eindeutigkeit gleich und entgegengesetzt sind (vgl. Terwilliger 1962). Im Unterschied zur nominalen Antonymie fordert Mordkoff (1963, 1965), daß polare Gegensatzpaare auch funktional antonym sein müßten, derart, daß sich SD-Beurteilungen der ‚bipolaren‘ Adjektive (Konzepte) als symmetrische und äquidistante Profile zum Nullpunkt - zumindest für die einzelnen Skalen, wenn schon nicht über die Skalen hinweg - darstellen ließen. Die mittels Hotellings T^2 geprüften Abweichungen von dieser Bedingung erwiesen sich in einer Vielzahl von Fällen als signifikant, nominell antonyme Adjektivpaare also keineswegs gesichert auch als funktional gegensätzlich. Bei systematischer Variation der konzeptspezifischen Information bilanzierten Malmstrom & French (1963) im Hinblick auf die Symmetrie von evaluativen SD-Skalen günstiger: Die Polarität der Urteile korrespondierte eng mit der Polarität der auf E-Skalen gegebenen Konzept-Information.

Auch Andersons (1970) Ähnlichkeits-Ratings von 12 EPA-Adjektiven bestätigen - trotz bemerkenswerter Variation bei Polaritäten und Personen - im wesentlichen die Angemessenheit der Bipolaritäts-Annahme, und zwar sowohl im Hinblick auf die Größe der Distanzen zwischen bipolaren Adjektiven, wie auch auf die Äquidistanz vom Ursprung.

Green & Goldfried (1965) argumentieren, daß die vorgegebene bipolare Etikettierung der Rating-Skalen die Überprüfung einer dementsprechenden Bipolarität der Urteilsdimensionen nicht erlaubt. Die Bipolarität der EPA-Struktur sei durch die Anordnung der Adjektivskalen erzwungen. Sie selbst präsentierten ihren Vpn die Antonyme jeweils einzeln (unipolar) zur Beurteilung von Konzepten und konstatierten erhebliche Abweichungen von funktionaler Bipolarität bei EPA-Skalen, die sich in Null- oder positiven Korrelationen zwischen ‚gegensätzlichen‘ Adjektiven äußern. Demgegenüber hatte Ertel (1964) aufgrund einer ähnlichen Vorgehensweise der Korrelierung von Daten unipolarer Skalen diejenigen als faktoriell gegensätzlich gefunden, die bereits intuitiv als Gegensätze angesehen worden waren. Nach Green & Goldfried (1965) sind alle Tendenzen von gegensätzlichen Adjektiven, Pole einer einzigen Skala zu bilden, konzept- oder konzeptklassenabhängig. Ausgehend von der größeren Anfälligkeit von unipolaren Rating-Skalen für Zustimmungstendenzen berichtet Bentler (1969) für Adjektiv-Zuordnungen zu den Polen der EPA-Dimensionen ähnliche Korrelationskoeffizienten wie Green & Goldfried

(1965), die sich jedoch dramatisch zur Übereinstimmung mit der Bipolaritäts-Annahme verändern, wenn Zustimmungstendenz als die Gesamtzahl der Adjektive, die über alle sechs Pole verwendet wurden, kontrolliert wurde. Bipolarität von Skalen, die die Bipolarität der EPA-Dimensionen repräsentieren sollen, haben Atwood & Falkenberg (1971) für unipolare SD-Ratings nach Auspartialisierung von Zustimmungstendenz als die Summe aller Skalenmarkierungen über alle Konzepte für jeden Beurteiler weniger deutlich und allenfalls für konkrete Konzepte (Makkaroni, Akkordeon, Automobile) im Unterschied zu abstrakten (Glück, Notwendigkeit, Theorie) gefunden.

Gegenüber Erklärungen der beobachtbaren Abweichungen von der funktionalen Bipolaritätsannahme (i.S. von Skalensymmetrie) durch Person- oder Konzept-Merkmale hat Gilpin (1973) lexikalische Markierungseffekte, eine strukturelle Asymmetrie der Skalen, als Bedingung verantwortlich gemacht. Er bezieht sich auf das Prinzip lexikalischer Markierung (Clark **1969**). Danach werden solche Adjektive als „unmarkiert“ bezeichnet, die in zweifacher Weise verwendet werden: ‚nominal‘, insofern sie das gesamte Kontinuum einer bipolaren Skala bezeichnen (z.B. ‚günstig‘ das Günstigkeitskontinuum von extrem günstig bis extrem ungünstig) und ‚kontrastiv‘, insofern sie eine Spezifizierung im Hinblick auf einen Standard oder Kontrast implizieren (z.B. günstig im Unterschied zu ungünstig). Markierte Adjektive weisen demgegenüber nur eine, nämlich kontrastive Bedeutung auf (z.B. ungünstig). Auf die Frage wie gut, interessant oder wichtig z.B. ein neues Produkt ist, wird die Antwort einer nominalen oder kontrastiven Verwendung dieser Adjektive entsprechen; der Befragte ist auf die kontrastive Bedeutung festgelegt, wenn die Frage lautet wie schlecht, uninteressant oder unwichtig das Produkt ist. Differenzen der Urteilsmaße auf unmarkierten und markierten unipolaren Skalen zu denen auf ihren bipolaren Skalen waren für die markierten Adjektive signifikant größer, mithin einen Effekt der Asymmetrie erzeugend. Aufgrund des Umstandes, daß unmarkierte Adjektive überwiegend evaluativ positive Bedeutung haben und markierte negative (Hamilton & Deese **1971**) ist dieser Befund auch für eine entsprechend alternative Interpretation offen.

Kaplan (1972) kritisiert, daß aufgrund der Definition der Mittelkategorie durch Osgood et al. (1957, 29 und 83) Indifferenz und Ambivalenz konfundiert sind. Osgood habe zwar die Gegensätzlichkeit der Adjektivpaare als Repräsentanten der reziprok-antagonistischen Tendenzen konzipiert, den Vpn würde jedoch die Mittelkategorie als ‚weder-noch‘ und als ‚sowohl-als auch‘-Kategorie erläutert. Auf diese Weise würden ambivalente Urteilstendenzen, statt als Markierungen auf beiden Seiten der Skala aufzutreten (was mit der Konzeption reziprok-antagonistischer Prozesse nicht vereinbar ist), in der Neutralkategorie aufgefangen. (Dazu ist zu bemerken, daß Osgood diesen Sachverhalt nicht nur gesehen, sondern theoretisch postuliert hat, s.O., S. 157). Mit Bezug auf die evaluativen SD-Skalen schlägt Kaplan eine Trennung

der Antonyme als orthogonale (liking-disliking) Komponenten vor, die jeweils durch unipolare Ratings (0 - 3 bzw. -3 - 0) zu erfassen sind und - zusätzlich zu den üblichen bipolaren SD-Daten - eine unabhängige Einschätzung des Ausmaßes von Ambivalenz erlauben sollen.

Die vorliegenden Befunde zur Bipolaritäts-Annahme sind überaus uneinheitlich. Dieser Sachverhalt läßt sich kaum zureichend mit den verwendeten, unterschiedlichen Operationalisierungen der Bipolarität begründen, da auch ähnliche Operationalisierungen zu unterschiedlichen Ergebnissen führen und verschiedene Operationalisierungen zu ähnlichen Ergebnissen. Vielmehr scheinen Varianten der Kombination von Skalen mit Konzepten dafür verantwortlich zu sein. Der Gegensatz von ‚gut‘ ist nicht invariant ‚schlecht‘, sondern u.U. auch ‚böse‘ (Brandt 1972). Da die Angemessenheit des polar-gegensätzlichen Schemas nicht grundsätzlich in Zweifel steht, sondern konkrete Formulierungen die Zweifel an der Geltung der Bipolaritäts-Annahme begründen, bedarf die Wahl von Antonymen in jedem Falle einer empirischen Begründung, und zwar unter Berücksichtigung der spezifischen Urteiler- und Konzeptpopulation.

3.2 Intervallgleichheit

Bei der Verarbeitung von SD-Daten wird - sowohl von Osgood & Cie, wie von anderen SD-Raumkonstrukteuren und -Anwendern - in aller Regel davon ausgegangen, daß die Antwortkategorien der (meist 7-stufigen) Skalen das bipolare Kontinuum nach gleichen Intervallbreiten aufteilen.

Soweit Adverbien die Kategorien auf den Beurteilungs-Skalen definieren, können für die von Osgood et al. (1957, 1975) verwendeten (slightly, quite, extremely) etwa gleiche Intensitätszuwächse angenommen werden (Cliff 1959; vgl. auch Howe 1962, 1966 a, 1966 b). Vergleiche der durch die numerische Kodierung bestimmten Kategoriengrenzen mit solchen, die aufgrund einer Skalierung nach dem Gesetz des kategorialen Urteils bestimmt sind, sind von Messick (1957), Revenstorff (1973a) und Fuchs (1974) durchgeführt worden. Diese Arbeiten unterscheiden sich im Hinblick auf die Anzahl der zur Beurteilung gegebenen Kategorien (7,10 bzw. 9) und die Skalen- (und Konzept-) Auswahlen (Standard Osgood, Standard Ertel, bzw. konzeptspezifisch). Die Ergebnisse stimmen in den wesentlichen Punkten überein:

- Die Abweichungen von den vorgegebenen gleicherscheinenenden Intervallen sind beträchtlich: Die Intervallbreiten verengen sich zur Skalenmitte hin und erscheinen besonders auf der positiv markierten Seite der Skalen gedehnt. Bei Revenstorff (Skala ohne Mittelpunkt) und Fuchs erscheint auch die mittlere Kategorie gedehnt. Fuchs berichtet - im Unterschied zu Messick - über erhebliche Variabilität zwischen den Skalen.

- Die Regression der skalierten auf die angenommenen Intervallgrenzen ist generell linear: Die entsprechenden Korrelationskoeffizienten liegen bei allen Autoren um .97 und darüber.

Nach diesen Befunden erscheinen insbesondere die auf dem euklidischen Distanzmodell basierenden D-Maße (Osgood & Suci 1952; Cronbach & Gleser 1953; Osgood et al. 1957) problematisch, da sie Intervallgleichheit auf den Skalen und über die Skalen voraussetzen (vgl. Diehl & Schäfer 1975).

Die von Revenstorff (1973a, 125) geäußerte Erwartung, „daß man in einer Faktorenanalyse der skalierten Durchschnittsprofile sehr ähnliche Aussagen über den konnotativen Raum gewinnen würde, wie bei der Faktorisierung der unskalierten Profile“ wird durch einen entsprechenden Befund von Fuchs (1974) auch empirisch belegt. Allerdings geht in die Erwartung die Voraussetzung ein, daß die verwendeten Skalen in einem gemeinsamen Nullpunkt symmetrisch sind.

3.3 Nullpunktlage

Die Arbeiten von Messick (1957) und Fuchs (1974) enthalten auch Hinweise, daß der subjektive Nullpunkt der Skalen nicht mit dem Skalenmittelpunkt zusammenfällt, sondern leicht zur Seite der ‚positiven‘ Skalenmarkierung verschoben ist. Dem entsprechen im Zusammenhang mit der Bipolaritäts-Annahme berichtete Befunde der Skalen-Asymmetrie. Anderson (1970) hat die Winkel bestimmt, die die bipolaren Adjektiv-Punkte mit dem Nullpunkt bilden: Sie sind nach seinen Befunden befriedigend zum Nullpunkt distanz-symmetrisch. Die Gemeinsamkeit des Nullpunktes setzt im Ursprung Cosinus-Werte von -1.00 voraus, d.h. die Geraden zwischen den polaren Punkten bilden im Nullpunkt einen Winkel von 180°. Aufgrund der beobachteten Abweichung konstatiert Anderson, daß die Ergebnisse allenfalls als ein schwacher Beleg für die Annahme eines gemeinsamen Nullpunktes angesehen werden können.

Wenn Dawes (1977) demonstriert, daß die korrelations-analytische Behandlung von gemittelten Rating-Urteilen, einschließlich solcher auf einer SD-Skala, bei Größenschätzungen zu Ergebnissen führt, die Messungen auf der Basis einer Repräsentationstechnik entsprechen, so mag dies als Beleg dafür gelten, daß Ratingmaße nützlich verwendet werden können. Für die SD-Technik ist dieser Anwendungsfall aber zumindest uncharakteristisch; der schon von Heise (1969) beklagte Mangel an Forschung zu Problemen der Metrik von SD-Skalen wird durch derartige Befunde kaum geringer.

4. Wahl von SD-Skalen zur Exploration von Bedeutungs- Räumen: Konstruktion von Semantischen Differentialen

Bei der Verwendung eines SDS als Instrument zur Erfassung der Bedeutung von Konzepten kommt der Antwort auf die beiden folgenden Fragen erhebliche Bedeutung zu:

- Wird die EPA-Struktur als ein angemessenes Bezugssystem zur Bedeutungsdifferenzierung akzeptiert?
- Lassen sich die relevanten Bedeutungsdimensionen durch einen Standardsatz von SD-Skalen repräsentieren?

Wenn - wie bei einem Großteil der SD-Verwendungen - beide Fragen bejaht werden, wird der Forscher von aufwendigen, technischen Konstruktionsarbeiten entlastet, und die Wahl von SD-Skalen wird zu einer Kompositions-Aufgabe: Das Material bilden die Ladungsmuster der Skalen aufgrund vorliegender Faktorenlösungen. Als semantische ‚Raum-Fähre‘ ist das SD - u.U. bei Modifikation der äußeren Verkleidung - jederzeit wiederverwendungsfähig.

Angesichts der Beobachtungen von Konzept-Skalen-Interaktionseffekten und ihrer Erklärung durch Bedeutungsverschiebungen bei den Skalen empfehlen Osgood et al. (1957, 78 ff) beim Einsatz des SDS zur Analyse der Bedeutung spezifischer Konzepte die Wahl der Skalen auch unter dem Gesichtspunkt der Relevanz, der semantischen Stabilität und der Bipolarität im Hinblick auf die untersuchten Konzepte vorzunehmen. Durch diese Zusatzkriterien wird allerdings nicht nur die Generalität der Indikatorfunktion von SD-Skalen, sondern auch die Verfügbarkeit einer Standardliste von Beurteilungsmerkmalen, die zur Lösung der Kompositions-Aufgabe herangezogen werden können, in Frage gestellt: Die Erfüllung derartiger Kriterien kann nicht impressionistisch, sondern nur auf der Grundlage systematisch-empirischer Prüfung gewährleistet werden. Vorliegende Merkmalslisten bieten dafür keine hinreichende Grundlage.

Aufgrund seiner - abweichenden - Einschätzung der Konzept-Skalen-Interaktionsproblematik konnte Ertel (1964, 1965a, 1965 b) eine Standardform des SDS vorschlagen, die den Anspruch genereller Verwendbarkeit erhebt, mit der Einschränkung allerdings, daß diejenigen Skalen aus der Merkmalsliste eliminiert werden müßten, die die Bedeutung eines Konzeptes ‚im eigentlichen Sinne‘, d.h. nicht-metaphorisch, denotativ spezifizieren. (Diese Einschränkung steht im Zusammenhang mit Ertels Bemühen, die Art der mit dem Eindrucksdifferential zu erfassenden Bedeutungsprozesse als „emotionale“ zu präzisieren.) Abgesehen von der unregelmäßigen Handhabung derartiger Ausschlüsse zieht Ertel auf diese Weise offenbar nur eine restriktive Konsequenz aus dem von Osgood - wie vorher erörtert - mit ‚denotativer Kontamina-

tion' bezeichneten Problem, das zur Erklärung von Konzept-Skalen-Interaktionseffekten herangezogen wird und zur Abkehr vom Ideal eines universell - über alle Konzepte und Personen - verwendbaren Standard-Instruments führte.

Osgood et al. (1957, 76) betonen, daß es unangemessen sei, das SD als ein bestimmtes Instrument (als eine Art Test) zu apostrophieren. Vielmehr handelt es sich um eine verallgemeinerbare Forschungstechnik, die jedem Forschungszweck anzupassen sei. Diese vorsichtige Einschätzung wird von den Autoren allerdings nicht durchgängig vertreten - so z.B. anders bei der Diskussion des SD als verallgemeinerter Einstellungs-Skala (p. 189) - und von Anwendern häufig unbeachtet gelassen. Durch den Bezug auf spezifische Untersuchungsbedingungen wird im übrigen das EPA-System nicht obsolet. Es ist auch nicht als eine problematische oder überflüssige Zugabe anzusehen, wie es - explizit - bei Darnell (1970) bzw. - implizit - bei Hofstätter (1955, 1959) geschieht. Vielmehr wird man - Miron (1969) zustimmend, wenn er behauptet, daß die EPA-Struktur des semantischen Raumes viel häufiger unter verschiedenen Bedingungen repliziert worden ist als die meisten anderen nicht unmittelbar beobachtbaren „Fakten“ unserer Wissenschaft - diese auch als Strukturmerkmal von SD-Skalen anzusehen haben. Abgesehen von den Einschränkungen bei der Reproduktion der EPA-Struktur für spezifische Konzepte/Konzeptklassen, die methodisch bedingt sein können, bedeutet die Akzeptierung der EPA-Struktur weder, daß diese Urteilsdimensionen alle Konzepte/Konzeptklassen in gleicher Weise charakterisieren, noch, daß sie bei allen Konzepten gleichermaßen durch identische Merkmale optimal repräsentierbar sind. Der EPA-Raum ist aufgrund des gewählten Analyseansatzes ein ‚Durchschnitts‘-Raum, der diejenigen Bedeutungsaspekte repräsentiert, die Beurteilern im Hinblick auf Konzepte gemeinsam sind. Die korrespondierenden Bedeutungsprozesse, die diesen Raum konstituieren, manifestieren sich im Urteilsverhalten als „gut“ reactions‘ (Osgood 1971, 37), d.h. abgesehen von der ‚ausgemittelten‘ Variation, die durch sprachliche Äußerungen von Personen im Hinblick auf Konzepte verursacht wird. Darüber hinaus stehen auch Abweichungen bei der Dimensionalität für Personen(gruppen) und Konzepte/Konzeptklassen nicht notwendig im Widerspruch zur ‚allgemeinen‘ EPA-Struktur; es ist bereits darauf hingewiesen worden, daß derartige spezifische Bedeutungsaspekte (bei Personen und für Konzepte) als Ergänzungen zur EPA-Durchschnittsstruktur verstanden werden können.

Im folgenden werden die wichtigsten Schritte der Konstruktion eines SD skizziert, wobei - je nach Forschungsintention - sowohl der Orientierung an der allgemeinen EPA-Struktur wie auch an spezifischen Bedeutungsstrukturen Raum gegeben wird. Das Verfahren berücksichtigt weitgehend die Vorgehensweise, die Osgood et al. (1975) zur Begründung der pankulturellen EPA-Struktur gewählt haben.

4.1 Merkmals-Relevanz

Bizarre Kombinationen von Konzepten und Skalen hatten schon in einer der ersten Würdigungen des ‚Measurement of Meaning‘ die Skepsis des Rezensenten hervorgerufen (Brown 1958: ‚Is a boulder sweet or sour?‘). Osgood et al. (1957, 78f) hatten dieses Problem durchaus erkannt und die Relevanz der verwendeten Skalen für die zu beurteilenden Konzepte als Kriterium formuliert. Für den Fall der Verwendung irrelevanter Skalen hatten sie eine Tendenz zu uncharakteristischen ‚Neutral‘-Urteilen festgestellt. Oetting (1967) hat diesen Sachverhalt aufgrund entsprechender Beobachtungen bestätigt. An dieser Stelle ist auch ein Befund von Mitsos (1961) zu erwähnen, der für die jeweils drei persönlich bedeutsamsten von je sieben typischen EPA-Skalen größere Distanzen zwischen (7) Konzepten und größere Distanzen der Konzeptpunkte vom Bedeutungsnullpunkt fand als für die übrigen Skalen. Reduzierte Variabilität der Konzeptbeurteilungen, die in unmittelbarem Zusammenhang mit Validitätseinbußen steht, wird von Orlik (1965) für subjektiv ‚sachlich nicht einschlägige‘ und von Grimm et al. (1973) für ‚allgemeine‘ gegenüber ‚inhaltsorientierten‘ SD-Skalen berichtet. Mikula & Schulter (1970, 383) bekräftigen diesen Befund und spezifizieren, daß „annähernd 45% der Gesamtvarianz der Einstufungen durch die ‚Geeignetheit‘ der verwendeten Polarität determiniert sind“. Dabei wird eine Tendenz zur Verwendung extremerer Skalenkategorien bei verbal begabteren Versuchspersonen festgestellt. Diese Befunde entsprechen der „meaningful-polarization“-Hypothese von O’Donovan (1965), wonach Reaktionen auf bedeutsame stimuli polarisiert werden, während Urteile auf bedeutungslose stimuli in Richtung auf die Indifferenz-Kategorie tendieren. Nur geringe Ähnlichkeit/Unterschiede zwischen den Konzeptbeurteilungen anhand subjektiv sachlich einschlägiger und weniger einschlägiger SD-Skalen berichtet dagegen Schick (1968). Eher wirkte sich objektive (d.h. inferenz-statistisch definierte) Trennschärfe auf die Höhe der einzelnen Konzeptähnlichkeiten aus; die Struktur der Ähnlichkeitsbeziehung zwischen den Konzepten für die beiden nach der ‚objektiven Trennschärfe‘ verschiedenen Skalensätze war allerdings wieder sehr ähnlich. Daß für die jeweils fünf in den Analysen von Osgood et al. (1957) am höchsten ladenden EPA-Indikatoren Relevanz für die Beurteilung anderer als der von diesen Autoren berücksichtigten Konzepte keineswegs gewährleistet ist, belegt ein Befund von Carter et al. (1969). Den Beurteilern war Gelegenheit gegeben, bei allen Skalen-Konzeptkombinationen (bei selbstgewählten Skalen-Antonymen) eine ‚wouldn’t use‘-Kategorie zu markieren. Von dieser Möglichkeit wurde in insgesamt 44% der Fälle, bei einigen Kombinationen zu mehr als 90%, Gebrauch gemacht.

Auf eine formale Bedingung der Angemessenheit von Konzept-Skalenkombinationen weisen Smith & Nichols (1973) hin: Konzepte wie Skalen sollten im Hinblick auf eher ‚intensionale‘ oder eher ‚extensionale‘ Bedeutung unterschieden werden. Durch diese Termini, die der philosophischen Sprachtradi-

tion entlehnt sind (Inhalt/Umfang), wird von diesen Autoren eher subjektive, ausschließlich konnotative Bedeutung abgegrenzt von objektiver, sowohl denotative wie konnotative Aspekte einschließende Bedeutung (z.B. idealistisch - realistisch bzw. sauber - schmutzig). Die faktorielle Instabilität als Ausdruck von Konzept-Skalen-Interaktionseffekten war reduziert, wenn Konzepte und Skalen der gleichen Bedeutungsart kombiniert waren.

In einer weiteren Gruppe von Untersuchungen wird gefragt, ob die mit erheblichem Aufwand verbundene Auswahl ‚inhaltsorientierter‘ oder ‚konzeptspezifischer‘ Skalen aufgrund der resultierenden Ergebnisse gerechtfertigt ist. Grimm et al. (1973) vertreten aufgrund ihrer Befunde die Ansicht, daß ein enger Problembezug der Skalen sich über die höhere Diskriminationsleistung der Skalen in einer deutlichen Validitätssteigerung auswirkt. Die unzureichende Begründung des verwendeten Validitätskriteriums in dieser Untersuchung mindert allerdings den Wert dieser Interpretation. Flade (1968) hält eine konzeptspezifisch zusammengestellte Merkmalsliste (Franke 1976; vgl. dazu Franke & Bortz 1972; Bortz 1972) einer allgemeinen, unspezifischen (Hofstätter 1971) gegenüber deshalb nicht für „geeigneter“, weil in beiden Fällen je drei Faktoren einen etwa gleichen Anteil der Konzeptvarianz erklären. Allerdings kann bezweifelt werden, daß die Spezifität der Merkmalsliste von Franke für die untersuchten Konzepte als adäquat gelten kann. Vor allem erscheint aber die Höhe des auf eine bestimmte Faktorenlösung entfallenden Varianzanteils als ‚Effizienz‘- und Präferenz-Kriterium fragwürdig, insbesondere da die Interpretation dieser Faktoren für die beiden Listen verschieden und ein Validitätskriterium nicht verfügbar ist.

Techniken, bei denen die Geeignetheit von Skalen in einer vorgeschalteten Erhebung durch Beurteiler eingestuft wird (Mills 1970; vgl. auch Mitsos 1961, Schick 1968, Mikula & Schulter 1970) setzen voraus, daß das relevante Beurteilungsrepertoire bereits bekannt ist. Sie sind deshalb für die Begründung der SD-Skalen nur sehr eingeschränkt tauglich.

Ein interessanter Ansatz, der es ermöglichen würde, ‚konzeptadäquat‘ (Bergler 1975) individuelle Bedeutungsstrukturen zu explorieren, ist von Micko (1962; vgl. auch Triandis 1959 a, 1959 b, 1960) in Anlehnung an Kellys ‚Role-Construct-Repertory-Test‘ vorgeschlagen worden: Die Personen werden gebeten, aus je drei aller Konzepte das gemeinsame Merkmal der beiden ähnlichsten und das Unterscheidungsmerkmal zum dritten Konzept zu benennen. Auf diese Weise wird für jede Person eine Liste von Beurteilungsmerkmalen gefunden, die zu einem individuellen SD zusammengestellt werden können. Zwar könnten die so gefundenen Merkmale auch entsprechend dem von Osgood und Mitarbeiter beim Cross-Cultural-Projekt angewendeten Verfahren (s.o.) behandelt werden, ein individualisierendes Vorgehen würde aber einen angemessenen Zugang zum Problem der Beurteilung interindividueller Differenzen im Hinblick auf Bedeutungsstrukturen ermöglichen; die berichteten

Lösungen von MDS-Analysen und dreimodalen Faktorenanalysen berücksichtigen zwar auch Personvarianz, setzen aber die Angemessenheit der allgemeinen Beurteilungsmerkmale (Skalen) für das Urteilsverhalten der untersuchten Personen voraus. In der Einstellungsforschung sind inzwischen Techniken erprobt, die die Analyse individueller Strukturen, sowie deren Aggregation zu Strukturtypen erlauben (Feger 1974, 1975). Eine solche Vorgehensweise ist aber angesichts des erheblichen Aufwandes der individuellen Erhebungen für die meisten der typischen SD-Anwendungsfälle nicht geeignet.

Das konventionelle, von Osgood et al. (1975; vgl. auch Fuchs & Schäfer 1972) entwickelte Verfahren, das im Bezugssystem des Durchschnitt-Bedeutungsraumes personengruppen- und konzept(klassen)-spezifischer Variation Rechnung tragen kann, dürfte für die meisten Fragestellungen zu brauchbaren Lösungen führen.

Für eine nach einem Repräsentativitätskriterium bestimmte Stichprobe von Konzepten aus der Population der zu untersuchenden, oder - soweit möglich - für alle zu untersuchenden Konzepte, werden (adjektivische) Qualifikatoren gesucht. Osgood und Mitarbeiter verwendeten das Frageschema „(Konzept) ist -“ und „(Das) - (Konzept)“. (Diese Erhebung sollte bei Personen durchgeführt werden, die die Personpopulation repräsentieren, die für die spätere Untersuchung in Betracht genommen ist.)

Die erhaltenden Adjektive werden sodann nach drei Kriterien geordnet: ‚salience‘ (Verwendungshäufigkeit über alle Konzepte), ‚diversity‘ (Zahl der verschiedenen Konzepte, für die die Adjektive verwendet wurden), ‚independence‘ (Ausmaß der Korreliertheit über die Konzepte). Die beiden ersten Kriterien können kombiniert als ‚productivity‘ durch Shannons H-Maß indiziert werden.

Die Adaptation lautet:

$$\text{index } H_j = - \sum_i p_{ij} \log_2 p_{ij} \quad (\text{productivity})$$

wobei i das Konzept- und j das Adjektiv-System bezeichnet, p_{ij} und $p_j(i)$ die Wahrscheinlichkeit des Auftretens eines Adjektivs bei allen bzw. den einzelnen Konzepten:

$$p_{ij} = f_{ij}/N_T \text{ und } p_j(i) = f_{ij}/N_j$$

Bezogen auf die absoluten Häufigkeiten läßt sich für die einzelnen Adjektive auch schreiben:

$$H = 1/N_T \left[\sum_i (f_{ij} \log_2 f_{ij}) - f_j \log_2 f_j \right]$$

H steigt mit der Gesamthäufigkeit eines Adjektivs und mit der Häufigkeit der Konzepte, für die es genannt wurde an. H wird = Null, wenn ein Adjektiv nur

bei einem einzigen Konzept verwendet wurde, unabhängig von der Häufigkeit der Verwendung. Das Maximum von H würde erreicht, wenn alle Personen für alle Konzepte dasselbe Adjektiv nennen würden.

Zur Reduzierung semantischer Redundanz verwenden Osgood und Mitarbeiter die Phi-Statistik als Index der Unabhängigkeit („Quasi-Synonymität“) von Merkmalen. Für jedes Adjektiv werden die Fälle von gemeinsamem und nicht-gemeinsamem Vorkommen (und Nicht-Vorkommen) mit jedem anderen, in der Rangordnung der H-Maße folgenden Adjektiv bei allen Konzepten gezählt und die Summen in die Berechnungsformel eingesetzt. Zur Vermeidung von Typ I-Fehlern (Ausschluß wegen angenommener Gleichheit trotz vorhandener Unterschiedlichkeit) wird eine hohe Signifikanzgrenze als Selektionskriterium verwendet.

Es kann zweifelhaft erscheinen, daß ein derartiges Maß stochastischer (Un-)Abhängigkeit semantische Synonymität angemessen operationalisiert. Überdies ist nicht geklärt, welche Auswirkungen diese ‚Säuberung‘ auf die weiteren Konstruktionsschritte, die Auswahlmöglichkeit von bipolaren Skalen und die Analyse der dimensional Struktur des Urteilsverhaltens hat. Eine so bestimmte Unabhängigkeit der Merkmale ist u.E. problematisch und entbehrlich.

4.2 Merkmals-Polarität

Die in Abschnitt 3.1 erörterten Argumente und Befunde lassen es notwendig erscheinen, die Bipolarität von SD-Skalen empirisch-systematisch und nicht bloß intuitiv zu begründen.

Osgood et al. (1975) haben Antonyme in jeder Sprache/Kultur-Gruppe durch jeweils ca. 10 kompetente Sprecher dieser Gruppen erhoben. Da aufgrund der einschlägigen Untersuchungen konzept- und personspezifische Variationen zu erwarten sind, erscheint für Forschung, die sich nicht unmittelbar auf die Analyse einer allgemeinen Bedeutungsstruktur bezieht, die Berücksichtigung entsprechender Besonderheiten nicht unwesentlich. Bei einer Stichprobe von Beurteilern aus der vorgesehenen Population von Untersuchungspersonen sollten demnach Antonyme für die einzelnen Adjektive im Hinblick auf die zu beurteilenden Konzepte erhoben werden: ‚Das Gegenteil von einem (Adjektiv) (Konzept) ist ein _____ (Konzept)‘.

Es gibt bislang weder systematische noch konventionelle Kriterien, die das Maß der noch akzeptablen Beurteiler-Nichtübereinstimmung spezifizieren. Fuchs & Schäfer (1972) weisen auf einen bei eindeutigen Gegensätzen zu erwartenden Sprung in der Häufigkeitskurve hin. Wenn zwei, jeweils relativ häufig verwendete Antonyme auftreten, dürfte es zweckmäßig sein, zunächst

beide beizubehalten, da sie möglicherweise verschiedene Urteilskontinua repräsentieren. In jedem Falle sollte überprüft werden, ob sich Abweichungen auf wenige Konzepte konzentrieren; ggf. ist zu erwägen, solche Konzepte als untypisch zu eliminieren.

4.3 Dimensionale Repräsentativität

Es ist festgestellt worden, daß die Repräsentation von Beurteilungs-Dimensionen ein charakteristisches Merkmal der SD-Technik ist. Mit der EPA-Struktur wird dafür ein allgemeines Bezugssystem zur Bedeutungs-differenzierung bereitgestellt. Soweit das Forschungsinteresse darauf gerichtet ist, die Bedeutung sehr verschiedenartiger Konzepte für Beurteiler vergleichbar zu machen, die nicht näher spezifiziert sind als durch ihre Zugehörigkeit zu einer Sprach/Kultur-Gemeinschaft, ist das Angebot, dafür den EPA-Raum zu wählen, konkurrenzlos. Die Reproduktion dieser Struktur durch Dimensionsanalysen von SD-Urteilen ist, wie mehrfach betont, von einer entsprechend allgemeinen, breit gefächerten Konzeptauswahl, die bereits im Prozeß der Merkmalsfindung zugrunde gelegt wird, abhängig. Wichtiges Material für die von Osgood et al. (1975) untersuchten Sprach/Kultur-Gruppen findet sich hierzu in dem der Veröffentlichung beigelegten ‚Semantischen Atlas‘.

Je spezifischer die untersuchte Konzeptklasse und Personengruppe ist, um so wichtiger wird die Frage, ob die EPA-Struktur als Vergleichsstandard dienen soll und kann, d.h. ob die aufgrund der Konstruktionsarbeit erhältliche spezifische Information ausgeschöpft werden oder eine Anpassung an die wohlbegründete, wenngleich unspezifische Struktur versucht werden soll. Durch eine Konzeptauswahl, die - ggf. mit Hilfe von Atlas-Daten - die Kombinationsmöglichkeiten der Oktanten des SD-Raumes abdeckt, durch Ergänzung der Skalenliste um EPA-Markierskalen, sowie durch entsprechende Rotation der Faktorenstruktur wird die letztere Lösung begünstigt. Heise (1969) ist der Auffassung, daß in einem solchen Falle mindestens 40 Konzepte (je 5 pro Oktanten) verwendet werden sollten, schon um die Skalenkorrelationen auf der Basis von Konzept-Mittelwerten bestimmen zu können. Wenn eine geringere Anzahl von Konzepten zugrunde gelegt wird, sollten die Skalenkorrelationen über $m \times n$, d.h. über alle Personen bei allen Konzepten berechnet und zusätzlich auftretende Faktoren ignoriert werden.

Ein unbedingtes Festhalten an der Verbindlichkeit der EPA-Struktur ist allerdings nicht begründbar. Es entspricht auch nicht der Konzeption Osgoods, der immer wieder betont hat, daß EPA zwar die zentralen Dimensionen affektiver Bedeutung von Zeichen repräsentiert, den Bedeutungsraum aber keineswegs im Hinblick auf Person- und Konzept-Variation erschöpfend beschreibt.

Die hier beschriebene, personengruppen- und konzeptklassen-spezifische Auswahl von SD-Skalen steht der Identifizierung von EPA-Dimensionen nicht im Wege. Sie erlaubt zwar keinen Vergleich und keine Typisierung individueller konzeptklassen-spezifischer Bedeutungsstrukturen, aber die Exploration weiterer und/oder konzeptspezifischer Bedeutungsdimensionen für bestimmte Personengruppen, vorausgesetzt, daß die untersuchte Konzeptklasse in der Konstruktionsphase hinreichend repräsentiert war.

4.4 Variationen der Präsentationsweise

4.4.1 Reihenfolge der Konzept-Skalenkombination

In der üblichen Form der Anwendung erhalten die Beurteiler Antwortbögen, auf denen die Skalen unterhalb der Nennung des jeweils zu beurteilenden Konzeptes in dimensional gemischter Reihenfolge und balancierter Polung der Bewertungsrichtung aufgeführt sind. In den früheren Arbeiten haben Osgood und Mitarbeiter eine Variante verwendet, bei der das Urteilskonzept zu jeder Skala neu festgelegt wird, wodurch eine Permutation der Konzept-Skalenkombinationen ermöglicht wird. Durch die Wahl dieser Form sollte die Wahrscheinlichkeit des Auftretens von Halo-Effekten verringert werden. Andererseits ist nicht auszuschließen, daß die Bedeutung der Konzepte durch die unterschiedlichen Kontextbedingungen stärker variiert. Im direkten Vergleich der Ergebnisse beider Verfahrensweisen fanden Osgood et al. (1957, 82) keine nennenswerten Unterschiede zwischen den Skalenmittelwerten. Die von ihnen des weiteren berichtete Resistenz von SD-Urteilen gegenüber Kontext-Ankereffekten wird von Sommer (1965) bekräftigt.

Mögliche kontextbedingte Fehlervarianz aufgrund einer Standardreihenfolge der Konzeptbeurteilungen wird häufig durch die technisch einfache Variation der Konzeptabfolge zu reduzieren versucht. Kane (1969) hat ein Computer-Programm entwickelt, mit dem sowohl die Reihenfolge der Konzepte, wie auch die Reihenfolge und Polung der Skalen systematisch permutiert werden können; in einer weiteren Arbeit (Kane 1971) berichtet er, daß eine Standardabfolge gegenüber verschiedenen Anordnungsvariationen weder im Hinblick auf die Faktorenstruktur und Faktoren-Scores, noch hinsichtlich der Markierungskonsistenz bei benachbarten Skalen zu unterschiedlichen Ergebnissen führt.

Osgood et al. (1975, 118 f) überprüften Ermüdungseffekte, soweit sie sich in einer geringeren oder gesteigerten Polarisierung der Skalenurteile auswirkten. Für umgekehrte Konzeptreihenfolgen wurden keine systematischen Effekte entdeckt (wobei allerdings Durchschnittswerte über die Personen zugrunde liegen).

4.4.2 Verankerung der Skalen

Die Uneinheitlichkeit der im Zusammenhang der Bipolarität von SD-Skalen berichteten Befunde kann - wie hier vorgeschlagen - zur Forderung einer systematischen Begründung der Antonyme von Merkmalen führen, die in bipolarer Anordnung als SD-Skalen verwendet werden. Wenn das bipolare Schema grundsätzlich in Frage gestellt wird, wird - allerdings nicht ohne Konsequenzen für die Konzeption der bedeutungsspezifischen Vermittlungsprozesse als reziprok-antagonistisch - die Verwendung unipolarer Skalen in Erwägung gezogen (z.B. Green & Goldfried 1965; Kaplan 1972).

Aus nicht näher beschriebenen Gründen haben Vidali (1973) und Vidali & Holeway (1975) eine Abwandlung der Green & Goldfried-Technik als ‚Stapel-Skalen‘ vorgeschlagen. Dabei wird die Skalenmitte durch ein Adjektiv gekennzeichnet, von dem die Ziffern 1 bis 3 auf- (+) und absteigen (-). Die Feststellung, daß die Ergebnisse insgesamt nicht signifikant verschieden von jenen ausfallen, die mit Hilfe von bipolaren SD-Skalen zu erhalten sind, kann allerdings nicht als eine hinreichende Begründung für den Vorzug oder eine ‚alternative‘ Verwendung von ‚Stapel-Skalen‘ akzeptiert werden. Im übrigen dürfte diese Präsentationsweise auch empfindlich für Effekte aufgrund ‚lexikalischer Markierung‘ sein, wonach die einzelnen Adjektive Urteilskontinua in unterschiedlicher Weise repräsentieren (vgl. S. 186).

Es werden auch Abweichungen vom adjektivischen Modus der Urteilsmerkmale vertreten. Ertel (1965a; vgl. auch Fuchs 1973) zieht die substantivische Form vor, weil er vermutet, daß diese die Beurteilung eher als Ähnlichkeitsvergleich fördert und nicht als Anheften von Attributen auffassen läßt. Er verspricht sich auch eine geringere Tendenz zu Verzerrungen im Sinne sozialer Erwünschtheit. Der Vergleich der Urteile auf einer adjektivischen und einer substantivischen Liste läßt gewisse Unterschiede zwischen diesen in der erwarteten Richtung erkennen (Ertel 1965 b).

Eine andere Abwandlung von der adjektivischen Skalenetikettierung wird von Mindak (196 1) vorgeschlagen und in der Markt- und Meinungsforschung häufiger verwendet: Die Beschreibungsmerkmale werden unabhängig von einer grammatikalischen Regel erhoben und verwendet, z.B. ‚really modern - sort of old-fashioned‘. Es scheint, daß auf diese Weise dem person- und konzeptspezifischen Urteilsverhalten sehr weitgehend Rechnung getragen werden kann, um den Preis allerdings, daß die Bestimmung verbindlicher Antonyme und bereits insoweit die Begründung eindimensionaler Kontinua erschwert ist.

4.4.3 Zahl der Antwortkategorien

Für die im Hinblick auf Rating-Skalen häufig diskutierte Frage nach der optimalen Anzahl der Abstufungskategorien (vgl. Miller 1956; McKelvie 1978) hat Gulliksen (1958) in bezug auf die SD-Technik die Forderung formuliert, 20- oder 30-stufige Skalen statt der von Osgood verwendeten sieben-stufigen Skala zu verwenden. Er begründet diese mit den von Osgood et al. (1957) berichteten hohen Übereinstimmungen im Urteilsverhalten bei Retests, die zeigten, daß die verlangte Diskriminationsleistung nicht zu einer Verteilung der Meßwerte führte, die die Bestimmung des Standardmeßfehlers erlaubt.

Die Zahl der tatsächlich verwendeten Skalenkategorien in SD-Untersuchungen schwankt nicht unerheblich um die 7, wobei kaum Skalen mit weniger als 5 und mehr als 10 Kategorien benutzt werden. Schönplflug (1972) hat die Auswirkungen dieser Variation für die Anzahlen von 3 bis 10 Kategorien systematisch untersucht. Für 15 Merkmale der Ertel-Liste (in adjektivischer Form), die zur Beurteilung von 48 Konzepten verwendet wurden, ergaben sich gleiche, dreidimensionale Bedeutungsräume mit gleicher Einlagerung der Konzepte, unabhängig von der kategorialen Differenziertheit der Skala. Die durchschnittlichen Einstufungen waren von der Kategorienzahl nicht abhängig, Urteile von Skalen verschiedener Kategorienzahl korrelierten alle nahezu perfekt.

Nach McKelvies (1978) Befunden aus einer nicht-SD-spezifischen Vergleichsstudie ist eine relativ geringe Zahl von Stufen empfehlenswert: Die Versuchspersonen operierten bei kontinuierlichen Skalen mit 5 oder 6 Stufen; 5-stufige Skalen wiesen die höchsten Reliabilitäten auf; bei stärkerer Abstufung (9 - 12 Stufen) ließen sich keine psychometrischen Vorteile belegen; bei weniger als 5 Stufen zeigte sich ein Verlust an Diskriminationsfähigkeit und Validität. ‚Die magische Zahl 7, plus oder minus 2‘ (Miller 1956), ob sie nun in der Kapazität menschlicher Informationsverarbeitung eine Grundlage hat oder nicht, dürfte als Maßgabe für die Differenzierung von SD-Skalen eine relevante Größe darstellen; dabei dürfte die Frage des Vorzugs einer verbalen Kodierung, wie von Osgood und Mitarbeitern verwendet, gegenüber einer numerischen nicht von Wichtigkeit sein.

Dagegen ist die Verwendung und Definition der Mittelkategorie problematisch. Im Unterschied zu Kaplans (1972) Kritik der mehrfachen Bedeutung dieser Kategorie, erscheint die Verwendung der ‚Neutral‘-Kategorie für den Fall ambivalenter Urteilstendenzen begründet (vgl. Osgood 1977, 16). Forthman (1973) kritisiert allerdings zu Recht, daß diese Kategorie in der Definition als Neutral-, Ambivalenz- und als Irrelevanz-Kategorie eine Art Müllschluckerfunktion zu erfüllen hat. Er isoliert die Ambivalenzfunktion der Mittelkategorie und findet bei Nichtberücksichtigung von Urteilen i. S. der Irrelevanz- und (leider auch gleichzeitig) der Neutral-Instruktion Abweichun-

gen vom Ladungsmuster gemäß der EPA-Struktur, deren Replikation bei Verwendung der Original-Instruktion gelingt.

Die Notwendigkeit einer Irrelevanz-Kategorie wird um so geringer sein, je konzept- und beurteilerspezifischer SD-Indikatoren ausgewählt worden sind. Wenn aufgrund sehr heterogener Beurteiler- und/oder Konzeptstichproben Konzept-Skalenkombinationen zu vermuten sind, die als irrelevant angesehen werden, dürfte eine durch Instruktion angebotene Möglichkeit der Streichung allerdings erwägenswert sein.

Es sollte schließlich noch erwähnt werden, daß Anwender der SD-Technik der Instruktion häufig eine nur geringe Bedeutung beimessen. Abgesehen von der eben genannten Einschränkung ist das von Osgood et al. (1957, 82ff) gegebene Beispiel nachahmenswert.

4.5 Varianten der Technik

Neben den Anwendungen der SD-Technik in verschiedenen Bereichen der Semantik des (verbalen) Urteilsverhaltens (z.B. ‚Persönlichkeits-Differential‘ Kuusinen 1969, Warr & Haycock 1970, Revenstorff 1971, Tzeng 1975, 1977; ‚Angst-Differential‘ Alexander & Husek 1962; ‚Verhaltens-Differential‘ Triandis 1971; ‚Stereotyp-Differential‘ Gardner et al. 1972; ‚Befindlichkeits-Differential‘ Baumann & Dittrich 1972; ‚face differential‘ Hurwitz et al. 1975) ist die Adaptation der SD-Technik zur Bedeutungsanalyse auf der Grundlage sprachfreier Bedeutungsreaktionen, die Entwicklung von ‚Grafischen Differentialen‘ erwähnenswert. Die mit Befunden aus der Analyse von Synästhesien begründete Auffassung, daß EPA die gemeinsamen affektiven Bedeutungsanteile bei Zeichen verschiedener Modalität repräsentiert (Osgood et al. 1957, Osgood 1959, Elliott & Tannenbaum 1963, Osgood et al. 1975), läßt die Wahl nicht-sprachlicher Indikatoren zur Bedeutungsdifferenzierung möglich und gelegentlich wünschenswert erscheinen (Jakobovits 1969). Kontrastive Paare visueller Muster ließen insbes. E, weniger deutlich P und A identifizieren. Die Replikation der EPA-Struktur grafischer Zeichen neben anderen Dimensionen des Bedeutungsraumes (vgl. Bentler & LaVoie 1972a) gelang Bentler & LaVoie (1972 b).

Ein wichtiger Anwendungsbereich der SD-Technik liegt in der Einstellungsforschung. Grundlegend sind dafür die von Osgood et al. (1957) gegebene Begründung, daß die von traditionellen Einstellungsskalen erfaßten Urteilstkontinua sowohl begrifflich wie empirisch mit der E-Dimension der EPA-Struktur übereinstimmen, sowie die Überzeugung, die typischen E-Skalen könnten als ein Standardsatz von Indikatoren für beliebige Einstellungsobjekte verwendet werden. Die Konsequenzen aus der zuletzt genannten Auffassung

sind zwar ‚benutzerfreundlich‘; sie ist jedoch nicht konsistent zu der Auffassung, die hier und von den Urhebern der Technik an anderer Stelle zur Verwendung von SDs vertreten wird (s.o.). Ein Überblick über die Verwendung der SD-Technik in der Einstellungsforschung geben Heise (1970) und Schäfer (1975c).

Kriterien der Konstruktion von Semantischen Differentialen können dazu beitragen, die SD-Technik als ein Verfahren zur Analyse der Bedeutung von Zeichen zu begründen. Die Fortentwicklung der SD-Technik wird allerdings erheblich vom Erfolg abhängen, mit dem die Nützlichkeit von SD-Bedeutungsmaßen für die Erklärung von Verhalten - über Urteilsverhalten hinaus - belegt werden kann.

Literatur

- Adams, F. M. & Osgood, C. E. 1973. A cross-cultural study of the affective meanings of color. *Journal of Cross-Cultural Psychology*, 4, 135-156.
- Alexander, S. & Husek, T. R. 1962. The anxiety differential. Initial steps in the development of a measure of situational anxiety. *Educational and Psychological Measurement*, 22, 325-348.
- Aiken, E. G. 1965. Alternate forms of a semantic differential for measurement of changes in self-description. *Psychological Reports*, 16, 177-178.
- Allison, R. B. 1963a. A two-dimensional semantic differential. *Journal of Consulting Psychology*, 27, 18-23.
- Allison, R. B. 1963 b. Using adverbs as multipliers in semantic differentials. *Journal of Psychology*, 56, 115-117.
- Amsfeld, Elizabeth & Lambert, W. E. 1964. Evaluational reactions of bilingual and monolingual children to spoken languages. *Journal of Abnormal and Social Psychology*, 69, 89-97.
- Amster, Harriet. 1964. Evaluative judgment and recall in incidental learning. *Journal of Verbal Learning and Verbal Behavior*, 3, 466-473.
- Anderson, A. B. 1970. Structure of semantic space. In: Borgatta, E. & Bohrnstedt, G. (eds) *Sociological methodology*. San Francisco: Jossey-Bass, 308-325.
- Anisfeld, M. & Lambert, W. E. 1966. When are pleasant words learned faster than unpleasant words? *Journal of Verbal Learning and Verbal Behavior*, 5, 132-141.
- Anisfeld, M., Bogo, N. & Lambert, W. E. 1962. Evaluation reactions to accented English speech. *Journal of Abnormal and Social Psychology*, 65, 223-231.
- Anisfeld, M., Munoz, S. R. & Lambert, W. E. 1963. The structure and dynamics of the ethnic attitudes of Jewish adolescents. *Journal of Abnormal and Social Psychology*, 66, 31-36.

- Arnold, J. B. 1971. A multidimensional scaling study of semantic distance. *Journal of Experimental Psychology (Monograph)*, 90, 349-372.
- Arthur, A. Z. 1965. Clinical use of the semantic differential. *Journal of Clinical Psychology*, 21, 337-338.
- Arthur, A. Z. 1966. Response bias in the semantic differential. *British Journal of Social and Clinical Psychology*, 5, 103-107.
- Atwood, J. T. & Falkenberg, S. D. 1971. The bipolarity of semantic space as related to concept complexity. *Journal of Psychology*, 79, 119-122.
- Barclay, A. & Thumin, F. J. 1963. A modified semantic differential approach to attitudinal assessment. *Journal of Clinical Psychology*, 19, 376.
- Barnard, J. W. 1966. The effects of anxiety on connotative meaning. *Child Development*, 37, 461-472.
- Barrett, G. V. & Otis, J. L. 1967. The semantic differential as a measure of changes in meaning in education and vocational counseling. *Psychological Reports*, 20, 335-338.
- Baumann, U. & Dittrich, A. 1972. Überprüfung der deutschen Version eines Polaritätenprofils zur Erfassung der Befindlichkeit. *Zeitschrift für klinische Psychologie*, 4, 335-350.
- Baxter, J. C. 1962. Mediated generalization as a function of semantic differential performance. *American Journal of Psychology*, 75, 66.
- Benel, Denise, C. R. & Benel, R. A. 1976. A further note on sex differences on the semantic differential. *British Journal of Social and Clinical Psychology*, 15, 437-439.
- Bentler, P. M. 1969. Semantic space ist (approximately) bipolar. *Journal of Psychology*, 71, 33-40.
- Bentler, P. M. & LaVoie, A. L. 1972a. An extension of semantic space. *Journal of Verbal Learning and Verbal Behavior*, 11, 38-47.
- Bentler, P. M. & LaVoie, A. L. 1972b. A nonverbal semantic differential. *Journal of Verbal Learning and Verbal Behavior*, 11, 491-496.
- Bergler, R. (ed) 1975. *Das Eindrucksdifferential. Theorie und Technik*. Bern: Huber.
- Bettinghaus, E. P. 1963. Cognitive balance and the development of meaning. *Journal of Communication*, 8, 94-105.
- Birch, D. & Erickson, M. 1958. Phonetic symbolism with respect to three dimensions from the semantic differential. *Journal of General Psychology*, 58, 291-297.
- Berlyne, D. E. & Peckham, S. 1966. The semantic differential and other measures of reaction to visual complexity. *Canadian Journal of Psychology*, 20, 125-135.
- Black, H. K. 1975. Semantic differential ratings of impoverished Stimuli: A replication. *Bulletin of the Psychonomic Society*, 5, 81-83.
- Block, J. 1958. An unprofitable application of the semantic differential. *Journal of Consulting Psychology*, 22, 235-236.

- Bobbitt, R. G. & Beck, R. C. **1971**. Semantic differential judgments of single and multiple conditioned Stimuli with an aversive delay conditioning paradigm. *Journal of Experimental Psychology*, 89, 398-402.
- Bokander, I. 1966. Semantic description of complex and meaningful stimulus material. *Perceptual and Motor Skills*, 22, 201-202.
- Bortz, J. 1972. Beiträge zur Anwendung der Psychologie auf den Städtebau. II. Erkundungsexperiment zur Beziehung zwischen Fassadengestaltung und ihrer Wirkung auf den Betrachter. *Zeitschrift für experimentelle und angewandte Psychologie*, **19**, 226-281.
- Bousfield, W. A. 1961. The problem of meaning in verbal learning. In: Cofer, C. N. (ed) *Verbal learning and verbal behavior*. New York: McGraw-Hill, 81-91.
- Brandt, L. W. 1972. Questions concerning some assumptions underlying the semantic differential. *Psychologische Beiträge*, 14, 61-67.
- Brandt, L. W. 1978. Messung eines Maßstabs: Empirische Untersuchung des Semantischen Differentials (SD). *Probleme und Ergebnisse der Psychologie*, 66, 71-74.
- Brewer, W. F. & Lichtenstein, E. H. 1974. Memory for marked semantic features versus memory for meaning. *Journal of Verbal Learning and Verbal Behavior*, 13, 172-180.
- Brinton, J. E. 1961. Deriving an attitude scale from semantic differential data. *Public Opinion Quarterly*, 25, 289-295.
- Brod, Diane, Kernoff, Phyllis & Terwilliger, R. F. 1964. Anxiety and semantic differential responses. *Journal of Abnormal and Social Psychology*, 68, 570-574.
- Brown, R. 1958. Is a boulder sweet or sour? *Contemporary Psychology*, 3, 113-115.
- Burger, G. K. & Pickett, L. 1976. The California psychological inventory and the semantic differential dimensions. *Journal of General Psychology*, 94, 129-134.
- Burns, R. 1976. The concept-scale interaction problem: A trap for the unwary on the semantic differential. *Educational Studies*, 2, 121-127.
- Bynner, J. & Romney, D. 1972. A method for overcoming the problem of concept-scale interaction in semantic differential research. *British Journal of Psychology*, **63**, 229-234.
- Carroll, J. B. 1959. Review of „The measurement of meaning“. *Language*, 35, 58-77.
- Carroll, R. M. & Field, J. 1974. A comparison of the classification accuracy of profile similarity measures. *Multivariate Behavioral Research*, 9, 373-380.
- Carter, R. F., Ruggels, W. L. & Chaffee, S.H. 1969. The semantic differential in opinion measurement. *Public Opinion Quarterly*, 32, 666-674.
- Cassel, R. N. 1970. Development of a semantic differential to assess the attitude of secondary school and College students. *Journal of Experimental Education*, 39, **10-14**.
- Chen, Kathleen, C. 1977. Extremity of semantic differential ratings in deaf and hearing subjects. *Journal of General Psychology*, 96, 231-236.

- Clark, H. H. 1969. Linguistic processes in deductive reasoning. *Psychological Review*, **76**, 387-404.
- Clark, Virginia A. & Kerrick, J. S. 1967. A method of obtaining summary scores from semantic differential data. *Journal of Psychology*, **66**, 77-85.
- Clevenger, T., Lazier, G. A. & Clark, Margaret L. 1969. The influence of certain factors on response to the semantic differential. *Public Opinion Quarterly*, **32**, 675-679.
- Cliff, N. 1959. Adverbs as multipliers. *Psychological Review*, **66**, 27-44.
- Cowen, L. 1972. Anxiety, self-concept, and the semantic differential. *Journal of Psychology*, **80**, 65-68.
- Coyne, L. & Holzman, P. S. 1966. Three equivalent forms of a semantic differential inventory. *Educational and Psychological Measurement*, **26**, 665-674.
- Creelman, Marjorie B. 1966. The experimental investigation of meaning: A review of the literature. New York: Springer.
- Crockett, W. H. & Nidorf, L. J. 1967. Individual differences in responses to the semantic differential. *Journal of Social Psychology*, **73**, 211-218.
- Cronbach, L. J. & Gleser, Goldine C. 1953. Assessing similarity between profiles. *Psychological Bulletin*, **50**, 456-473.
- Darnell, D. K. 1966. Concept-scale interaction in the semantic differential. *Journal of Communication*, **16**, 106-115.
- Darnell, D. K. 1970. Semantic differentiation. In: Emmert, P. & Brooks, W. D. (eds) *Methods of research in communication*. Boston: Houghton Mifflin Company, 181-196.
- Dawes, R. M. 1977. Suppose we measured height with rating scales instead of rulers. *Applied Psychological Measurement*, **1**, 267-273.
- De Burger, R. A. & Donahoe, J. W. 1965. Relationship between the meanings of verbal Stimuli and their associative responses. *Journal of Verbal Learning and Verbal Behavior*, **4**, 25-31.
- Deese, J. 1964. The associative structure of some common English adjectives. *Journal of Verbal Learning and Verbal Behavior*, **3**, 347-357.
- Deese, J. 1965. The structure of associations in language and thought. Baltimore: John Hopkins Press.
- Denmark, Florence L., Shirk, Ethel J. & Riley, R. T. 1972. The effect of ethnic and social class variables on semantic differential Performance. *Journal of Social Psychology*, **86**, 3-9.
- Deutschmann, P. J. 1959. The semantic differential and public opinion research. *Public Opinion Quarterly*, **23**, 435.
- Diab, L. N. 1965. Studies in social attitudes. III: Attitude assessment through SD. *Journal of Social Psychology*, **67**, 303-314.
- Dicken, C. F. 1961. Connotative meaning as a determinant of stimulus generalization. *Psychological Monographs*, **75** (Whole No. 505).

- Diehl, Brigitte & Schäfer, B. 1975. Techniken der Datenanalyse beim Eindrucksdifferential. In: Bergler, R. (ed.) Das Eindrucksdifferential. Bern: Huber, 157-211.
- Dilts, Martha & Taylor, R. E. 1964. The semantic differential of color pyramid test instructions. *Perceptual and Motor Skills*, 19, 968-970.
- DiVesta, F. 1965. Developmental patterns in the use of modifiers as modes of conceptualization. *Child Development*, 36, 186-213.
- DiVesta, F. 1966a. A developmental study of the semantic structures of children. *Journal of Verbal Learning and Verbal Behavior*, 5, 249-259.
- DiVesta, F. 1966 b. A normative study of 220 concepts rated on the semantic differential by children in grades 2 through 7. *Journal of Genetic Psychology*, 109, 205-229.
- DiVesta, F. 1966 c. Norms for modifiers used by children in a restricted word-association task: Grades 2 through 6. *Psychological Reports*, 18, 65-66.
- DiVesta, F. & Dick, W. 1966. The test-retest reliability of children's ratings on the semantic differential. *Educational and Psychological Measurement*, 26, 605-616.
- DiVesta, F. & Stover, D. O. 1962. The semantic mediation of evaluative meaning. *Journal of Experimental Psychology*, 64, 467-475.
- Eisenman, R., Bernard, J. L. & Hannon, J. E. 1966. Benevolence, potency, and God: A semantic differential study of the Rorschach. *Perceptual and Motor Skills*, 22, 75-78.
- Eisenman, R. & Rappaport, Joan. 1967. Complexity preference and semantic differential ratings of complexity-simplicity and symmetry-asymmetry. *Psychonomic Science*, 7, 147-148.
- Eiser, J. R. 1971a. Enhancement of contrast in the absolute judgment of attitude Statements. *Journal of Personality and Social Psychology*, 17, 1-10.
- Eiser, J. R. 1971 b. Categorization, cognitive consistency and the concept of dimensional salience. *European Journal of Social Psychology*, 1, 435-454.
- Eiser, J. R. & Mower White, C. J. 1974. Evaluative consistency and social judgment. *Journal of Personality and Social Psychology*, 30, 349-359.
- Eiser, J. R. & Stroebe, W. 1972. Categorization and social judgment. London: Academic Press.
- Elliott, L. L. & Tannenbaum, P. H. 1963. Factor-structure of semantic differential responses to visual forms and prediction of factor-scores from structural characteristics of the Stimulus shapes. *American Journal of Psychology*, 76, 589-597.
- Endler, N. S. 1961. Changes in meaning during psychotherapy as measured by the semantic differential. *Journal of Counseling Psychology*, 8, 105-111.
- Ertel, S. 1964. Die emotionale Natur des „semantischen“ Raumes. *Psychologische Forschung*, 28, 1-32.
- Ertel, S. 1965 a. Standardisierung eines Eindrucksdifferentials. *Zeitschrift für experimentelle und angewandte Psychologie*, 12, 22-58.

- Ertel, S. 1965 b. Weitere Untersuchungen zur Standardisierung eines Eindrucksdifferentials. *Zeitschrift für experimentelle und angewandte Psychologie*, 12, 177-208.
- Ertel, S. 1969. *Psychophonetik: Untersuchungen über Lautsymbolik und Motivation*. Göttingen: Hogrefe.
- Ervin-Tripp, Susan M. & Slobin, D.I. 1966. Psycholinguistics. *Annual Review of Psychology*, 17, 435-474.
- Etkind, A. M. 1979. (Versuch einer theoretischen Interpretation des semantischen Differentials), *Voprosy Psichologii* 1, 17-27 (Orig. russ., engl. Übersetzung: *Soviet Psychology* 1980, Vol. 18, 3-20).
- Everett, A. V. 1973. Personality assessment at the individual level using the semantic differential. *Educational and Psychological Measurement*, 33, 837-844.
- Feger, H. 1974. Die Erfassung individueller Einstellungsstrukturen. *Zeitschrift für Sozialpsychologie*, 5, 242-254.
- Feger, H. 1975. Längsschnittliche Erfassung intraindividuellder Unterschiede bei Einstellungsstrukturen. In: Lehr, U. & Weinert, F. (eds) *Entwicklung und Persönlichkeit*. Stuttgart, 38-49.
- Feger, H. & Wiczorek, R. 1980. Multidimensionale Skalierung in der Einstellungsmessung. In: Petermann, F. (ed.) *Einstellungsmessung - Einstellungsforschung*. Göttingen, 153-174.
- Finley, J. R. & Staats, A. W. 1967. Evaluative meaning words as reinforcing stimuli. *Journal of Verbal Learning and Verbal Behavior*, 6, 193-197.
- Flade, Antje. 1978. Die Beurteilung umweltpsychologischer Konzepte mit einem konzeptspezifischen und einem universellen Semantischen Differential. *Zeitschrift für experimentelle und angewandte Psychologie*, 15, 367-378.
- Flavell, J. H. 1961a. Meaning and meaning similarity: I. A theoretical reassessment. *Journal of General Psychology*, 64, 307-319.
- Flavell, J. H. 1961 b. Meaning and meaning similarity: II. The semantic differential and co-occurrence as predictors of judged similarity in meaning. *Journal of General Psychology*, 64, 321-335.
- Flavell, J. H. & Johnson, Ann. 1961. Meaning and meaning similarity: III. Latency and number of similarities as predictors of judged similarity in meaning. *Journal of General Psychology*, 64, 337-348.
- Fodor, J. A. 1965. Could meaning be an r_m ? *Journal of Verbal Learning and Verbal Behavior*, 4, 73-81.
- Fodor, J. A. 1966. More about mediators: A reply to Berlyne and Osgood. *Journal of Verbal Learning and Verbal Behavior*, 5, 412-415.
- Ford, L. H. & Meisels, M. 1965. Social desirability and the semantic differential. *Educational and Psychological Measurement*, 24, 465-475.
- Forthman, J. H. 1973. The effect of a zero interval on semantic differential rotated factor loadings. *Journal of Psychology*, 84, 23-32.

- Franke, J. 1976. Die Erlebniswirkung von Wohnumgebungen. In: Kaminski, G. (ed.) *Umweltpsychologie*. Stuttgart, 134-143.
- Franke, J. & Bortz, J. 1972. Beiträge zur Anwendung der Psychologie auf den Städtebau. I. Vorüberlegungen und erste Erkundungsuntersuchungen zur Beziehung zwischen Siedlungsgestaltung und Erleben der Wohnumgebung. *Zeitschrift für experimentelle und angewandte Psychologie*, 19, 76-108.
- Friedman, C. J. & Gladden, J. W. 1964. Objective measurement of social role concepts via the semantic differential. *Psychological Reports*, 14, 239-247.
- Friedman, C. J., Johnson, C. A. & Fode, K. 1964. Subjects descriptions of selected TAT cards via the semantic differential. *Journal of Consulting Psychology*, 28, 317.
- Fuchs, A. 1973. Emotionale Bedeutung religiöser Konzepte. Methoden-kritische Untersuchungen mit einem semantischen Differential. Phil. Diss., Universität Bonn (Fotodruck).
- Fuchs, A. 1974. Untersuchungen zu metrischen Problemen der Technik der Bedeutungsdifferenzierung. *Archiv für Psychologie*, 126, 114-124.
- Fuchs, A. 1975a. Grundzüge einer Verhaltenstheorie der Bedeutung. In: Bergler, R. (ed.) *Das Eindrucksdifferential*. Bern: Huber, 33-68.
- Fuchs, A. 1975 b. Das Eindrucksdifferential als Instrument zur Erfassung emotionaler Bedeutungsprozesse. In: Bergler, R. (ed.) *Das Eindrucksdifferential*. Bern: Huber, 69-100.
- Fuchs, A. & Schäfer, B. 1972. Kriterien und Techniken der Merkmalsselektion bei der Konstruktion eines Eindrucksdifferentials. *Archiv für Psychologie*, **124**, **282-302**.
- Gärling, T. 1976. A multidimensional scaling and semantic differential technique study of the perception of environmental settings. *Scandinavian Journal of Psychology*, 17, 323-332.
- Gardner, R. C., Taylor, D. M. & Feenstra, H. J. 1970. Ethnic stereotypes: Attitudes or beliefs? *Canadian Journal of Psychology*, 24, 321-334.
- Gardner, R. C., Wonnacott, E. J. & Taylor, D. M. 1968. Ethnic stereotypes: A factor analytic investigation. *Canadian Journal of Psychology*, 22, 35-44.
- Gardner, R. C., Kirby, D. M., Gorospe, F. H. & Villamin, A. C. 1972. Ethnic stereotypes: An alternative assessment technique, the stereotype differential. *Journal of Social Psychology*, 87, 259-267.
- Gleser, Goldine C., Cronbach, L.J. & Rajaratnam, N. 1965. Generalizability of scores influenced by multiple sources of variance. *Psychometrika*, 30, 395-418.
- Gilpin, A. R. 1973. Lexical marking effects in the semantic differential. *Journal of Psychology*, 85, 277-285.
- Godfrey, R. R. & Natalicio, L. F. S. 1970. Evaluation on semantic differential equals abstraction plus error. *Psychological Reports*, 27, 467-473.
- Goldfried, M. R. 1962. On differences in meaning between normals and neurotics. *Psychological Reports*, 11, 183.

- Goldfried, M. R. 1963. The connotative meaning of some animal symbols for College students. *Journal of Projective Technique*, 27, 60-67.
- Green, P. E., Aheshwari, A. & Rao, V. R. 1969. Dimensional interpretation and configuration invariance in multidimensional scaling: An empirical study. *Multivariate Behavioral Research*, 4, 159-180.
- Grigg, A. E. 1959. Validity study of the semantic differential technique. *Journal of Clinical Psychology*, 15, 179-181.
- Green, R. F. & Goldfried, M. R. 1965. On the bipolarity of semantic space. *Psychological Monographs*, 79 (whole No. 599).
- Grimm, G., Lück, H. E. & Timaeus, E. 1973. Zur internen Validität von Polaritätsprofilen: Untersuchungen zum Standard-Polaritätsprofil von Ertel. *Zeitschrift für experimentelle und angewandte Psychologie*, 20, 547-571.
- Gulliksen, H. 1958. How to make meaning more meaningful. *Contemporary Psychology*, 3, 115-118.
- Hamilton, D. L. 1968. Personality attributes associated with extreme response style. *Psychological Bulletin*, 69, 192-203.
- Hamilton, H. W. & Deese, J. 1971. Does linguistic marking have a psychological correlate? *Journal of Verbal Learning and Verbal Behavior*, 10, 707-714.
- Harman, H. H. 1970. *Modern factor analysis*. Chicago: University of Chicago Press.
- Hastorf, A. H., Osgood, C. E. & Ono, H. 1966. The semantics of facial expressions and the prediction of the meanings of stereoscopically fused facial expressions. *Scandinavian Journal of Psychology*, 7, 179-188.
- Haygood, R. C. 1966. Use of semantic differential dimensions in concept learning. *Psychonomic Science*, 5, 305-306.
- Heaps, R. A. 1972. Use of the semantic differential technique in research: Some precautions. *Journal of Psychology*, 80, 121-125.
- Heise, D. R. 1965. Semantic differential profiles for 1,000 most-frequent words. *Psychological Monographs*, 79, No. 8.
- Heise, D. R. 1969. Some methodological issues in semantic differential research. *Psychological Bulletin*, 72, 406-422.
- Heise, D. R. 1970. The semantic differential and attitude research. In: Summers, G. F. (ed.) *Attitude measurement*. Chicago: Rand McNally, 235-253.
- Herrmann, T. 1962. Urteilsnuanciertheit und Fraktionierung des Bezugssystems. Eine Zweifaktorenhypothese des semantischen Differenzierens. *Psychologische Beiträge*, 7, 539-557.
- Heskin, K. J., Bolton, N. & Smith, F. V. 1973. Measuring the attitudes of prisoners by the semantic differential. *British Journal of Social and Clinical Psychology*, 12, 73-77.
- Hoar, J. R. & Meek, E. E. 1965. The semantic differential as a measure of subliminal message effects. *Journal of Psychology*, 60, 165-169.
- Hörmann, H. 1976. *Meinen und Verstehen*. Frankfurt/M.

- Hoffmann, Kristine. 1976. Die Messung des Erlebens von Wohnumgebungen - Zur Problematik der Reliabilitätsbestimmung. In: Kaminski G. (ed.) *Umweltpsychologie*. Stuttgart, 144-155.
- Hofstätter, P. R. 1955. über Ähnlichkeit. *Psyche*, 9, 54-80.
- Hofstätter, P. R. 1959. *Einführung in die Sozialpsychologie*. Stuttgart.
- Hofstätter, P. R. 1971. *Differentielle Psychologie*. Stuttgart.
- Hogg, J. 1969. A principal components analysis of semantic differential judgments of single colors and color pairs. *Journal of General Psychology*, 80, 129-140.
- Homzie, M. J. & Weimer, J. 1967. Connotative similarity and paired-associate learning. *Journal of Verbal Learning and Verbal Behavior*, 6, 136-138.
- Hopper, D. & Padden, D. 1965. Psychiatric roles and their meaning. *British Journal of Social and Clinical Psychology*, 4, 35-38.
- Horman, M. 1960. Implicit personality theories of clinicians as defined by semantic structures. *Journal of Consulting Psychology*, 24, 180-186.
- Horn, J. L. 1965. An empirical comparison of methods for estimating factor scores. *Educational and Psychological Measurement*, 25, 313-322.
- Hovland, C. I. & Sherif, M. 1952. Judgmental phenomena and scales of attitude measurement: Item displacement in Thurstone scales. *Journal of Abnormal and Social Psychology*, 47, 822-832.
- Howe, E. S. 1962. Probabilistic adverbial qualifications of adjectives. *Journal of Verbal Learning and Verbal Behavior*, 1, 225-242.
- Howe, E. S. 1964. Three-dimensional structure of ratings of exploratory responses shown by a semantic differential. *Psychological Reports*, 14, 187-196.
- Howe, E. S. 1965a. Further data concerning the dimensionality of ratings of the therapists' verbal exploratory behavior. *Journal of Consulting Psychology*, 29, 73-76.
- Howe, E. S. 1965b. Uncertainty and other correlates of Osgood's D₄. *Journal of Verbal Learning and Verbal Behavior*, 4, 498-509.
- Howe, E. S. & Pope, B. 1960. Multiple scaling of therapists' responses with a semantic differential. *American Psychologist*, 15, 415.
- Howe, E. S. 1966a. Associative structure of quantifiers. *Journal of Verbal Learning and Verbal Behavior*, 5, 156-162.
- Howe, E. S. 1966b. Verb tense, negatives, and other determinants of the intensity of evaluative meaning. *Journal of Verbal Learning and Verbal Behavior*, 5, 147-155.
- Hurwitz, D., Wiggins, Nancy Hirschberg & Jones, L. E. 1975. A semantic differential for facial attribution: The face differential. *Bulletin of the Psychonomic Society*, 6, 370-372.
- Husek, T. R. & Wittrock, M. C. 1962. The dimensions of attitudes toward teachers as measured by the semantic differential. *Journal of Educational Psychology*, 53, 209.
- Irle, M. 1975. *Lehrbuch der Sozialpsychologie*. Göttingen.

- Jakobovits, L. A. 1966a. Comparative psycholinguistics in the study of cultures. *International Journal of Psychology*, 1, 15-37.
- Jakobovits, L. A. 1966b. Mediation theory and the „single-stage“ S-R model: Different? *Psychological Review*, 73, 376-381.
- Jakobovits, L. A. 1969. The affect of Symbols: Towards the development of a cross-cultural graphic differential. *International Journal of Symbolology*, 1, 28-52.
- Jenkins, J. J. 1960. Degree of polarization and scores on the principal factors for concepts in the semantic atlas study. *American Journal of Psychology*, 73, 274-279.
- Jenkins, J. J., Russell, W. A. & Suci, G. J. 1958. An atlas of semantic profiles for 360 words. *American Journal of Psychology*, 71, 688-699.
- Jenkins, J. J., Russell, W. A. & Suci, G. J. 1959. A table of distances for the semantic atlas. *American Journal of Psychology*, 72, 623-625.
- Johnson, R. L. & Wall, D. D. 1969. Cluster analysis of semantic differential data. *Educational and Psychological Measurement*, 29, 769-780.
- Jones, J. M. 1969. Meanings of ‚most acceptable‘ and ‚most objectionable‘. *Psychological Reports*, 24, 915-921.
- Jones, J. M. 1970. Dimensions of meaning and attitude change. *Psychological Reports*, 26, 955-962.
- Jones, J. M. 1971. Attitudinal valence and semantic differential potency scales. *Psychological Reports*, 28, 991-994.
- Katz, M. 1965. Agreement on connotative meaning in marriage. *Family Process*, 4, 64-75.
- Kahneman, D. 1963. The semantic differential and the structure of inferences among attributes. *American Journal of Psychology*, 76, 554-567.
- Kane, R. B. 1969. Computer generation of semantic differential (SD) questionnaires. *Educational and Psychological Measurement*, 29, 191-192.
- Kane, R. B. 1971. Minimizing order effects in the semantic differential. *Educational and Psychological Measurement*, 31, 137-144.
- Kanungo, R. N. & Lambert, W. E. 1963. Semantic satiation and meaningfulness. *American Journal of Psychology*, 76, 421-428.
- Kaplan, K. J. 1972. On the ambivalence-indifference problem in attitude theory and measurement: A suggested modification of the semantic differential technique. *Psychological Bulletin*, 77, 361-372.
- Kashiwagi, S. 1965. Geometric vector orthogonal solution for the semantic differential scales of Sagara et al. *Psychological Reports*, 16, 914.
- Kaufman, Helen J. 1959. The semantic differential: A critical appraisal. *Public Opinion Quarterly*, 23, 437-438.
- Keil, C. & Keil, Angeliki. 1966. Musical meaning: A preliminary report. *Ethnomusicology*, 10, 153-173.

- Kelly, Jane A. & Levy, L. H. **1961**. The discriminability of concepts differentiated by means of the semantic differential. *Educational and Psychological Measurement*, **21**, 53-58.
- Kentler, H. 1959. Zur Problematik der Profilmethode. *Diagnostica*, 5, 5-18.
- Kilty, K. M. 1972. Attitudinal affect and behavioral intentions. *Journal of Social Psychology*, **86**, 251-256.
- King-Fun Li, Anita. 1966. The Cantonese semantic differential scales. *Journal of Education*, 23.
- Kirby, D. M. & Gardner, R. C. 1973. Ethnic stereotypes: Determinants in children and their parents. *Canadian Journal of Psychology*, 27, 127-143.
- Kjeldergaard, P. M. 1961. Attitudes toward newscasters as measured by the semantic differential: A descriptive case. *Journal of Applied Psychology*, 45, 35-40.
- Kjeldergaard, P. M. & Higa, M. 1962. Degree of polarization and the recognition value of words selected from the semantic atlas. *Psychological Reports*, **11**, 629-630.
- Klapper, J. T. 1959. The semantic differential: Its use and abuse. *Public Opinion Quarterly*, 23, 435-438.
- Klemmack, D. L. & Ballweg, J. A. 1973. Concept-scale interaction with the semantic differential technique. *Journal of Psychology*, 84, 345-352.
- Koen, F. 1962. Polarization, m, and emotionality in words. *Journal of Verbal Learning and Verbal Behavior*, **1**, 183-187.
- Komorita, S. S. & Bass, A. R. 1967. Attitude differentiation and evaluative scales of the semantic differential. *Journal of Personality and Social Psychology*, 6, 241-244.
- Korman, M. 1960. Implicit personality theories of clinicians as defined by semantic structures. *Journal of Consulting Psychology*, 24, 180-186.
- Kostić, D. & Das, Rhea S. 1971. Aspects of meaning revealed by the semantic differential technique. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 24, 55-75.
- Krause, Ingrid & Solle, R. 1973. über die faktorielle Struktur des Werte-Raumes bei französischen Studenten. *Zeitschrift für experimentelle und angewandte Psychologie*, 20, 210-239.
- Krieger, Margery H. 1963. A control for social desirability in a semantic differential. *British Journal of Social and Clinical Psychology*, 2, 94-103.
- Kroes, W. H. & Libby, W. L. 1971. Relative power of taxonomic, semantic differential, and sense impression categories for the organization of free recall. *Journal of General Psychology*, 85, 165-169.
- Kubinić, Cathleen M. & Farr, S. D. 1971. Concept-scale and concept-component interaction in the semantic differential. *Psychological Reports*, 28, 531-541.
- Kumata, Hideya. 1957. A factor analytic investigation of the generality of semantic structures across two selected cultures. Ph. D. Diss., University of Illinois.
- Kumata, Hideya & Schramm, W. 1956. A pilot study of cross-cultural methodology. *Public Opinion Quarterly*, 20, 229-238.

- Kuusinen, J. 1969. Affective and denotative structures of personality ratings. *Journal of Personality and Social Psychology*, 12, 181-188.
- Lambert, W. E. & Jakobovits, L. A. 1960. Verbal satiation and changes in the intensity of meaning. *Journal of Experimental Psychology*, 60, 376-383.
- Lana, R. E. & Pauling, F. J. 1965. Opinion change when the semantic differential is a pretest. *Psychological Reports*, 17, 730.
- Lane, Silvia T. M. 1973. Semantic differential scales for Portuguese speakers in Brazil. *International Journal of Psychology*, 8, 147-152.
- Lawson, E. D. 1971. Semantic differential analysis of men's first names. *Journal of Psychology*, 78, 229-240.
- Lawson, E. D. & Giles, H. 1973. British semantic differential responses on world powers. *European Journal of Social Psychology*, 3, 233-240.
- Lawson, E. D., Golden, G. H. & Chmura, Kathy J. 1972. Computer programs for the semantic differential. *Educational and Psychological Measurement*, 32, 779-784.
- Levin, J. 1965. Three-mode factor analysis. *Psychological Bulletin*, 64, 442-452.
- Levy, P. 1972. Concept-scale interaction in semantic differential research: Solutions in search of a problem. *British Journal of Psychology*, 63, 235-236.
- Light, C. S., Zax, M. & Gardiner, D. H. 1965. The relationship of age, sex, and intelligence level to extreme response style. *Journal of Personality and Social Psychology*, 2, 907-909.
- Lilli, W. 1975. Soziale Akzentuierung. Stuttgart.
- Litt, E. N. 1966. A factorial study of responses to abstract paintings. Unpubl. Master's Thesis. Univ. of Illinois (zit. nach Osgood et al. 1975).
- Lohr, J. M. 1976. Concurrent conditioning of evaluative meaning and imagery. *British Journal of Psychology*, 67, 353-358.
- Long, Barbara, H., Henderson, E. H. & Ziller, R. C. 1968. Self-ratings on the semantic differential: Content versus response set. *Child Development*, 39, 647-656.
- Lyle, J. 1960. Semantic differential scales for newspaper research. *Journalism Quarterly*, 37, 559-562, 646.
- Maclay, H. & Ware, E. E. 1961. Cross-cultural use of the semantic differential. *Behavioral Science*, 6, 185-190.
- Magnusson, D. & Ekman, G. 1970. A psychophysical approach to the study of personality traits. *Multivariate Behavioral Research*, 5, 255-274.
- Maguire, T. O. 1973. Semantic differential methodology for the structuring of attitudes. *American Educational Research Journal*, 10, 295-306.
- Malmstrom, E. J. & French, G. M. 1963. Scale-symmetry and the semantic differential. *American Journal of Psychology*, 76, 446-451.
- Madden, J. E. 1961. Semantic differential rating of self and of self-reported personal characteristics. *Journal of Consulting Psychology*, 25, 183.

- Maltz, H. E. 1963. Ontogenetic change in the meaning of concepts as measured by the semantic differential. *Child Development*, 34, 667-674.
- Manis, M. 1959. Assessing communication with the semantic differential. *American Journal of Psychology*, 72, 111-113.
- Markel, N. N. 1966. The validity of the semantic differential for psycholinguistic analysis. *Journal of Verbal Learning and Verbal Behavior*, 5, 348-350.
- Markel, N. N. & Meisels, M. 1964. Judging personality from voice quality. *Journal of Abnormal and Social Psychology*, 69, 458-463.
- Markel, N. N., Hunt, R. G. & Crapsi, L. A. 1966. The validity of the semantic differential for psycholinguistic analysis. *Journal of Verbal Learning and Verbal Behavior*, 5, 348-350.
- Marks, I. 1965. Patterns of meaning in psychiatric patients: Semantic differential responses in obsessives and psychopaths. London: Oxford University Press.
- Martinez, J. L. jr., Martinez, S. R., Olmedo, E. L. & Goldman, R. D. 1976. The semantic differential technique. A comparison of Chicano and Anglo high school students. *Journal of Cross-Cultural Psychology*, 7, 325-333.
- Maruyama, K. 1971. On the reliability of the data obtained by the semantic differential method. *Japanese Psychological Research*, 13, 51-59.
- Mayerberg, Cathleen K. & Bean, A. G. 1974. The structure of attitude toward quantitative concepts. *Multivariate Behavioral Research*, 9, 311-324.
- McCallon, E. L. & Brown, J. D. 1971. A semantic differential instrument for measuring attitude towards mathematics. *Journal of Experimental Education*, 39, 69-72.
- McKelvie, S. J. 1978. Graphic rating scales - How many categories? *British Journal of Psychology*, 69, 185-202.
- Meisels, M. & Ford, L. H. 1969. Social desirability response set and semantic differential evaluative judgments. *Journal of Social Psychology*, 78, 45-54.
- Messer, S., Jakobovits, L. A., Kanungo, R. & Lambert, W. E. 1964. Semantic satiation of words and numbers. *British Journal of Psychology*, 55, 155-163.
- Messick, S. J. 1957. Metric properties of the semantic differential. *Educational and Psychological Measurement*, 17, 200-206.
- Micko, H. C. 1962. Die Bestimmung subjektiver Ähnlichkeiten mit dem semantischen Differential. *Zeitschrift für experimentelle und angewandte Psychologie*, 9, 242-280.
- Mikula, G. & Schulter, G. 1970. Polaritätenauswahl, verbale Begabung und Einstufung im Polaritätenprofil. *Zeitschrift für experimentelle und angewandte Psychologie*, 17, 371-385.
- Miller, G. A. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.
- Miller, P. McC. 1974. A note on sex differences on the semantic differential. *British Journal of Social and Clinical Psychology*, 13, 33-36.
- Miller, S., James, Carol, Rytten, B., Tansill, R. & Thompson, C. 1971. Use of the

- semantic differential in the study of motivation. *Psychological Reports*, 29, **1279-1282**.
- Mills, D. H. 1970. Adjectives pertinent to psychotherapy for use with the semantic differential: An heuristic note. *Psychological Reports*, 26, 211-213.
- Mindak, W. A. 1961. Fitting the semantic differential to the marketing problem. *Journal of Marketing*, 25, 28-33.
- Miron, M. S. 1961. The influence of instruction modification upon test-retest reliabilities of the semantic differential. *Educational and Psychological Measurement*, 21, 883-893.
- Miron, M. S. 1969. What is it that is being differentiated by the semantic differential? *Journal of Personality and Social Psychology*, 12, 189-193.
- Miron, M. S. 1972. Universal semantic differential shell game. *Journal of Personality and Social Psychology*, 24, 313-320.
- Miron, M. S. & Osgood, C. E. 1966. Language behavior: The multivariate structure of qualification. In: Cattell, R. B. (ed.): *Handbook of multivariate experimental psychology*. Chicago: Rand McNally, 790-819.
- Mitsos, S. B. 1961. Personal constructs and the semantic differential. *Journal of Abnormal and Social Psychology*, 62, 433-434.
- Mogar, R. E. 1960. Three Versions of the F scale and Performance on the semantic differential. *Journal of Abnormal and Social Psychology*, 60, 262-265.
- Mordkoff, A. M. 1963. An empirical test of the functional antonymy of semantic differential scales. *Journal of Verbal Learning and Verbal Behavior*, 2, 504-508.
- Mordkoff, A. M. 1965. Functional vs nominal antonymy in semantic differential scales. *Psychological Reports*, 16, 691-692.
- Morimoto, H. 1957. A study of semantic relation by association method and the semantic differential. *Japanese Journal of Educational Psychology*, **4**, **131-137**.
- Moss, C. S. 1960. Current and projected status of semantic differential research. *Psychological Record*, 10, 47-54.
- Mueller, W. S. 1966. Anxiety level, inferred identification **and** response tendencies on a semantic differential. *Journal of Consulting Psychology*, 13, 149-152.
- Muthen, B., Pettersson, T., Olsson, U. & Stahlberg, G. 1977. Measuring religious attitudes using the semantic differential technique: An application of three-mode factor analysis. *Journal of the Scientific Study of Religion*, 16, 275-288.
- Neuringer, C. 1963. Effect of intellectual level and neuropsychiatric status on the diversity of intensity of semantic differential ratings. *Journal of Consulting Psychology*, 27, 280.
- Nickels, S. A. & Shaw, M. E. 1964. Saliency and two measures of attitude. *Psychological Reports*, 14, 273-274.
- Nordenstreng, K. 1968. A comparison between the semantic differential and similarity analysis in the measurement of musical experience. *Scandinavian Journal of Psychology*, 9, 89-96.

- Nordenstreng, K. 1969. Toward quantification of meaning. An evaluation of the semantic differential technique. *Annales Academiae Scientiarum Fennicae*, B 161, 2, Helsinki.
- Nordenstreng, K. 1970. Changes in the meaning of semantic differential scales: Measurement of subject-scale interaction effects. *Journal of Cross-Cultural Psychology*, 1, 217-327.
- Norman, W. T. 1959. Stability-characteristics of the semantic differential. *American Journal of Psychology*, 72, 581-584.
- O'Donovan, D. 1965. Rating extremity: Pathology or meaningfulness? *Psychological Review*, 72, 358-372.
- Oetting, E. R. 1964. Cross-cultural communication and the semantic differential: Research note. *Journal of Counseling Psychology*, 3, 292-293.
- Oetting, E. R. 1967. The effect of forcing response on the semantic differential. *Educational and Psychological Measurement*, 27, 699-702.
- Oles, H. J. 1973. Semantic differential for third through fifth grade students. *Psychological Reports*, 33, 24-26.
- Orlik, P. 1965. Eine Modellstudie zur Psychophysik des Polaritätsprofils. *Zeitschrift für experimentelle und angewandte Psychologie*, 12, 614-647.
- Orlik, P. 1967. Eine Technik zur erwartungsstreuenden Skalierung psychologischer Merkmalsräume aufgrund von Polaritätsprofilen. *Zeitschrift für experimentelle und angewandte Psychologie*, 14, 616-650.
- Osgood, C. E. 1952. The nature and measurement of meaning. *Psychological Bulletin*, **49**, 197-237.
- Osgood, C. E. 1959a. Semantic space revisited. *Word*, 15, 192-200.
- Osgood, C. E. 1959b. The cross-cultural generality of visual-verbal synesthetic tendencies. *Behavioral Science*, 5, 146-169.
- Osgood, C. E. 1962. Studies on the generality of affective meaning systems. *American Psychologist*, 17, 10-28.
- Osgood, C. E. 1964. Semantic differential technique in the comparative study of cultures. *American Anthropologist*, 66, 171-200.
- Osgood, C. E. 1965. Cross-cultural comparability in attitude measurement via multilingual semantic differentials. In: Steiner, I. & Fishbein, M. (eds) *Current studies in Social Psychology*. New York: Holt, Rinehart and Winston, 95-107.
- Osgood, C. E. 1966. Meaning cannot be r_m ? *Journal of Verbal Learning and Verbal Behavior*, 5, 402-407.
- Osgood, C. E. 1969. On the whys and wherefores of E, P, and A. *Journal of Personality and Social Psychology*, 12, 194-199.
- Osgood, C. E. 1971. Exploration in semantic space: A personal diary. *Journal of Social Issues*, 27, (4), 5-64.
- Osgood, C. E. & Luria, Z. 1954. A blind analysis of a case of triple personality using

- the semantic differential. *Journal of Abnormal and Social Psychology*, 49, **579-591**.
- Osgood, C. E. & Suci, G. 1952. A measure of relation determined by both mean difference and profile information. *Psychological Bulletin*, 49, **251-262**.
- Osgood, C. E. & Suci, G. 1955. Factor analysis of meaning. *Journal of Experimental Psychology*, 50, 325-338.
- Osgood, C. E., Suci, G. & Tannenbaum, P. 1957. *The measurement of meaning*. Urbana: University of Illinois Press.
- Osgood, C. E., May, W. H. & Miron, M. S. 1975. *Cross-cultural universals of affective meaning*. Urbana: University of Illinois Press.
- Osgood, C. E., Ware, E. E. & Morris, C. 1961. Analysis of the connotative meanings of a variety of human values as expressed by American College students. *Journal of Abnormal and Social Psychology*, 62, 62-73.
- Osipow, S. & Grooms, R. R. **1962**. On semantic differential resistance to response bias based on stimulus word position. *Psychological Reports*, **10**, **634**.
- Oyama, T., Tanaka, Y. & Chiba, Y. 1962. Affective dimensions of colors: A cross-cultural study. *Japanese Psychological Research*, **4**, **78-91**.
- Paivio, A. 1969. Mental imagery in associative learning and memory. *Psychological Review*, 76, 241-263.
- Paivio, A. **1971**. *Imagery and verbal processes*. New York: Holt, Rinehart and Winston.
- Peabody, D. 1962. Two components in bipolar scales: Direction and extremeness. *Psychological Review*, 69, 65-73.
- Piaggio, L. **1968**. Einstufungswahrscheinlichkeiten im semantischen Differential. *Zeitschrift für experimentelle und angewandte Psychologie*, **15**, **272-290**.
- Piaggio, L. 1969. Bestimmung der Skalenzahl im semantischen Differential. *Probleme und Ergebnisse der Psychologie*, 31, 5-13.
- Plutchik, R. 1967. The affective differential: Emotion profiles implied by diagnostic concepts. *Psychological Reports*, 20, 19-25.
- Prothro, E. T. & Keehn, J. D. 1957. Stereotypes and semantic space. *Journal of Social Psychology*, 45, 197-209.
- Presly, A. S. 1969. Concept-scale interaction in the semantic differential and its implications for factor scores. *British Journal of Psychology*, **60**, **109-113**.
- Priest, P. N. 1971. The influence of psychiatric status and sex on the semantic differential response style. *Personality: An International Journal*, 2, 9-14.
- Reece, M. W. 1964. Masculinity and femininity: A factor analytic study. *Psychological Reports*, 14, 123-139.
- Reed, T. R. 1972. Connotative meaning of social interaction concepts: An investigation of factor structure and the effects of imagined contexts. *Journal of Personality and Social Psychology*, 24, 306-312.

- Revenstorff, D. 1971. Persönlichkeitsbeurteilung auf dem Polaritätsprofil. *Archiv für Psychologie*, 123, 195-216.
- Revenstorff, D. 1973a. Zur Analyse des konnotativen Raumes. *Zeitschrift für experimentelle und angewandte Psychologie*, 20, 117-152.
- Revenstorff, D. 1973 b. über Profilähnlichkeit. *Archiv für Psychologie*, 125, 203-232.
- Roll, S. & Verinis, J. S. 1971. Stereotypes of scalp and facial hair as measured by the semantic differential. *Psychological Reports*, 28, 975-980.
- Rosenbaum, L. L. & McGinnies, E. 1969. A semantic differential analysis of concepts associated with the 1964 presidential election. *Journal of Social Psychology*, 78, 227-235.
- Rosenbaum, L. L., Rosenbaum, W. B. & McGinnies, E. 1971. Semantic differential factor structure stability across subject, concept, and time differences. *Multivariate Behavioral Research*, 6, 451-469.
- Ross, B. M. & Levy, N. 1960. A comparison of adjectival antonyms by simple card-pattern formation. *Journal of Psychology*, 49, 133-137.
- Ross, J. 1965. Change in the use of semantic differential with a change in context. *Journal of Verbal Learning and Verbal Behavior*, 4, 148-151.
- Rothman, A. I. 1968. The factor analysis of a science-related semantic differential instrument. *Journal of Educational Measurement*, 5, 145-149.
- Rydell, S. T. 1966. Tolerance of ambiguity and semantic differential ratings. *Psychological Reports*, 19, 1303-1312.
- Sagara, M., Yamamoto, K., Nishimura, H. & Akuto, H. 1961. A study on the semantic structure of Japanese language by the semantic differential method. *Japanese Psychological Research*, 3, 146-156.
- Schäfer, B. 1973. Die Messung der ‚Beurteilung von Völkern‘ mit Hilfe eines Eindrucksdifferentials. *Archiv für Psychologie*, 125, 29-38.
- Schäfer, B. 1975a. Konstruktion des Umfrage-Instrumentariums. In: Schweitzer, C. C. & Feger, H. (eds) *Das deutsch-polnische Konfliktverhältnis seit dem Zweiten Weltkrieg*. Boppard, 187-223.
- Schäfer, B. 1975 b. Konstruktion eines Eindrucksdifferentials zur Erfassung der ideologiespezifischen Bewertung politischer Schlüsselwörter. In: Bergler, R. (ed) *Das Eindrucksdifferential*. Bern: Huber, 139-155.
- Schäfer, B. 1975c. Das Eindrucksdifferential als Instrument zur Einstellungsmessung. In: Bergler, R. (ed) *Das Eindrucksdifferential*. Bern: Huber, 101-118.
- Schick, A. 1968. Der Einfluß systematischer Skalenauswahl auf die Verlaufsähnlichkeit von Polaritätsprofilen. *Zeitschrift für experimentelle und angewandte Psychologie*, 15, 146-160.
- Schlosberg, H. 1954. Three dimensions of emotion. *Psychological Review*, 61, 81-88.
- Schludermann, S. & Schludermann, E. 1969. Scale checking style as a function of age and sex in Indian and Hutterite children. *Journal of Psychology*, 72, 253-261.
- Schönpluf, W. 1972. Ein Problem bei der Arbeit mit Kategorienskalen: Welchen

- Einfluß hat die Zahl der Skalenkategorien? Zeitschrift für experimentelle und angewandte Psychologie, 19, 141-171.
- Schuh, A. J. 1966. Use of the semantic differential in a test of Super's vocational adjustment theory. Journal of Applied Psychology, 50, 516-522.
- Shell, S. A., O'Mally, J. M. & Johnsgard, K. W. 1964. The semantic differential and inferred identification. Psychological Reports, 14, 547-558.
- Shikiar, R., Fishbein, M. & Wiggins, Nancy. 1974. Individual differences in semantic space: A replication and extension. Multivariate Behavioral Research, 9, 201-209.
- Sines, J. O. 1962. An indication of specificity of denotative meaning based on the semantic differential. Journal of General Psychology, 67, 113-115.
- Singet-, R. D. 1961. A note on the use of the semantic differential as a predictive device in milieu therapy. Journal of Clinical Psychology, 17, 376-378.
- Smith, R. G. 1959. Development of a semantic differential for use with speech-related concepts. Speech Monographs, 26, 263-272.
- Smith, R. G. 1961. A semantic differential for theatre concepts. Speech Monographs, 28, 1-8.
- Smith, R. G. 1962. A semantic differential for speech correction concepts. Speech Monographs, 29, 32.
- Smith, R. G. 1963. Validation of a semantic differential. Speech Monographs, 30, 50-55.
- Smith, R. G. 1966. Semantic differential dimensions and form. Speech Monographs, 33, 17-22.
- Smith, R. G. & Nichols, H. J. 1973. Semantic differential stability as a function of meaning domain. Journal of Communication, 23, 64-73.
- Snider, J. G. 1962. Profiles of some stereotypes held by ninth-grade pupils. Alberta Journal of Educational Research, 8, 147-156.
- Snider, J. G. & Osgood, C. E. (eds) 1969. Semantic differential technique: A sourcebook. Chicago: Aldine.
- Snyder, F. 1967. An investigation of the invariance of the semantic differential across the subject mode. Unpubl. Master's Thesis, Univ. of Illinois (zit. nach Osgood et al. 1975).
- Snyder, F. W. & Wiggins, Nancy 1970. Affective meaning systems: A multivariate approach. Multivariate Behavioral Research, 5, 453-468.
- Solar-z, A. K. 1963. Perceived activity in semantic atlas words as indicated by a tapping response. Perceptual and Motor Skills, 16, 91-94.
- Sommer, R. 1965. Anchor-effects and the semantic differential. American Journal of Psychology, 78, 317-318.
- Springbett, B. M. 1960. The semantic differential and meaning in non-objective art. Perceptual and Motor Skills, 10, 231-240.

- Staats, A. W. 1968. Learning, language and cognition. New York: Holt Rinehart and Winston.
- Staats, A. W. & Staats, Carolyn K. 1957. Meaning established by classical conditioning. *Journal of Experimental Psychology*, 54, 74-80.
- Staats, A. W. & Staats, Carolyn K. 1959. Meaning and m: Separate but correlated. *Psychological Review*, 66, 136-144.
- Staats, A. W., Staats, Carolyn K. & Heard, W. G. 1961. Denotative meaning established by classical conditioning. *Journal of Experimental Psychology*, 61, 300-303.
- Stricker, G. 1963. The use of the semantic differential to predict voting behavior. *Journal of Social Psychology*, 59, 159-167.
- Stricker, G. & Zax, M. 1966. Intelligence and semantic differential discriminability. *Psychological Reports*, 18, 775-778.
- Stricker, G., Takahashi, S. & Zax, M. 1967. Semantic differential discriminability: A comparison of Japanese and American students. *Journal of Social Psychology*, 71, 23-25.
- Suci, G. J. 1960. A comparison of semantic structures in American Southwest culture groups. *Journal of Abnormal and Social Psychology*, 61, 25-30.
- Tajfel, H. 1959. Quantitative judgment in social perception. *British Journal of Psychology*, 50, 16-29.
- Tajfel, H. 1975. Soziales Kategorisieren. In: Moscovici, S. (ed.) *Forschungsgebiete der Sozialpsychologie* 1. Frankfurt/M., 345-380.
- Tanaka, Y. 1962. A Cross-cultural study of national stereotypes held by American and Japanese College graduate subjects. *Japanese Psychological Research*, 4, 65-78.
- Tanaka, Y. 1964. Studies on the measurement of meaning and the generality of affective meaning Systems: A review. *Japanese Psychological Research*, 8, 27-69.
- Tanaka, Y. & Osgood, C. E. 1965. Cross-culture, cross-concept, and cross-subject generality of affective meaning Systems. *Journal of Personality and Social Psychology*, 2, 143-153.
- Tanaka, Y., Oyama, T. & Osgood, C. E. 1963. A cross-culture and cross-concept study of the generality of semantic spaces. *Journal of Verbal Learning and Verbal Behavior*, 2, 392-405.
- Tannenbaum, P. H. 1959. Selected applications of the semantic differential. *Public Opinion Quarterly*, 23, 435-439.
- Taylor, C. L. & Haygood, R. C. 1968. Effects of degree of category separation on semantic concept identification. *Journal of Experimental Psychology*, 76, 356-359.
- Taylor, D. M. & Gardner, R. C. 1969. Ethnic stereotypes: Their effects on the perception of communicators of varying credibility. *Canadian Journal of Psychology*, 23, 161-173.

- Taylor, H. F. **1971**. Semantic differential factor scores as measures of attitude and perceived attitude. *Journal of Social Psychology*, 83, 229-234.
- Terwilliger, R. F. 1962. Free association patterns as a factor relating to semantic differential responses. *Journal of Abnormal and Social Psychology*, 65, 87-94.
- Triandis, H. C. 1959a. Categories of thought of managers, clerks, and workers about jobs and people in industry. *Journal of Applied Psychology*, 43, 338-344.
- Triandis, H. C. 1959b. Differential perception of certain jobs and people by managers, clerks and workers in industry. *Journal of Applied Psychology*, 43, 221-225.
- Triandis, H. C. 1960. A comparative factorial analysis of job semantic structures of managers and workers. *Journal of Applied Psychology*, 44, 297-302.
- Triandis, H. C. 1971. *Attitude and attitude change*. New York: Wiley.
- Triandis, H. C. & Osgood, C. E. 1958. A comparative factorial analysis of semantic structures in monolingual Greek and American College students. *Journal of Abnormal and Social Psychology*, 57, 187-196.
- Turvey, M. T. & Fertig, Joanne 1970. Polarity on the Semantic differential and release from proactive interference in short-term memory. *Journal of Verbal Learning and Verbal Behavior*, 9, 439-443.
- Turvey, M. T., Fertig, Joanne & Kravetz, S. 1969. Connotative classification and proactive interference in short-term memory. *Psychonomic Science*, 16, **223-224**.
- Tzeng, O. C. S. 1975a. Reliability and validity of semantic differential E-P-A markers for an American English representative sample. *Psychological Reports*, 37, 292.
- Tzeng, O. C. S. 1975b. Differentiation of affective and denotative meaning systems and their influence in personality ratings. *Journal of Personality and Social Psychology*, 32, 978-988.
- Tzeng, O. C. S. 1977. A quantitative method for separation of semantic subspaces. *Applied Psychological Measurement*, 1, 171-184.
- Tzeng, O. C. S. & May, W. H. 1975. More than E, P and A in semantic differential scales: An answer to questions raised by S.T.M. Lane. *International Journal of Psychology*, **10**, **101-117**.
- Tzeng, O. C. S. & Landis, D. 1978. Three-mode multidimensional scaling with points of view solutions. *Multivariate Behavioral Research*, 13, 181-213.
- Vidali, J. J. 1973. Single-anchor stapel scales versus double-anchor semantic differential scales. *Psychological Reports*, 33, 373-374.
- Vidali, J. J. 1976. Reliability of the semantic differential under practical conditions. *Psychological Reports*, 39, 583-586.
- Vidali, J. J. & Holeway, R. E. 1975. Stapel scales versus semantic differential scales: Further evidence. *Psychological Reports*, 36, 165-166.
- Voyce, C. D. & Jackson, D. N. 1977. An evaluation of a threshold theory for personality assessment. *Educational and Psychological Measurement*, 37, 383-408.
- Walker, Betty A. & Robinson, R. 1977. Using a semantic differential in predicting

- counselor-trainee' success in practicum. *Educational and Psychological Measurement*, 37, 971-975.
- Ware, E. E. 1958. Relationships of intelligence and sex to diversity of individual semantic meaning spaces. Unpubl. Doc. Diss., Univ. of Illinois (zit. nach Osgood et al. 1957; 1975).
- Warr, P. B. & Haycock, Valerie 1970. Scales for a British personality differential. *British Journal of Social and Clinical Psychology*, 9, 328-337.
- Washington, W. N. 1975. A methodology for response style analysis of the semantic differential index. *Journal of General Psychology*, 93, 289-294.
- Weigel, R. G., Weigel, Virginia M., Thornton, G. C. & Magnusson, F. 1975. Assessment of preferences among Company names by semantic differential and free association technique. *Psychological Reports*, 37, 1163-1166.
- Weinreich, U. 1958. Travels through semantic space. *Word*, 14, 346-366.
- Weinreich, U. 1959. A rejoinder (Osgood, C. E., Semantic space revisited.) *Word*, 15, 200-201.
- Weksel, W. & Hennes, J. D. 1965. Attitude intensity and the semantic differential. *Journal of Personality and Social Psychology*, 2, 91-94.
- Wells, W. D. & Smith, Georgianna. 1960. Four semantic rating scales compared. *Journal of Applied Psychology*, 44, 393-397.
- Wickens, D. D. 1970. Encoding categories of words: An empirical approach to meaning. *Psychological Review*, 77, 1-15.
- Wickens, D. D. & Clark, Sandra 1968. Osgood dimensions as an encoding class in short-term memory. *Journal of Experimental Psychology*, 78, 580-584.
- Wiggins, Nancy & Fishbein, M. 1969. Dimensions of semantic space: A problem of individual differences. In: Snider, J. G. & Osgood, C. E. (eds) *Semantic differential technique*. Chicago: Aldine, 183-193.
- Wilcox, R. C. 1966. Effects of context on semantic differential ratings. *Psychological Reports*, 18, 873-874.
- Wildman, R. W. II & Wildman, R. W. 1976. Note on application of the semantic differential to the electoral process. *Psychological Reports*, 38, 1185-1186.
- Williams, J. E. 1964. Connotations of color names among Negroes and Caucasians. *Perceptual and Motor Skills*, 18, 721-731.
- Williams, J. E. 1966. Connotations of racial concepts and color names. *Journal of Personality and Social Psychology*, 3, 531-540.
- Williams, J. E. & Carter, Dorothy J. 1967. Connotations of racial concepts and color names in Germany. *Journal of Social Psychology*, 72, 19-26.
- Williams, J. E., Morland, J. K. & Underwood, W. L. 1970. Connotations of color names in the United States, Europe, and Asia. *Journal of Social Psychology*, 82, 3-14.
- Williams, W. S. 1971. A semantic differential study of the meaning of personality test

- items to children from different socioeconomic groups. *Journal of Psychology*, **79**, 179-188.
- Williams, W. S. 1972. A study of the use of the semantic differential by fifth grade children from different socioeconomic groups. *Journal of Psychology*, **81**, 343-350.
- Winograd, E. 1966. Recognition memory and recall as a function of degree of polarization on the semantic differential. *Journal of Verbal Learning and Verbal Behavior*, **5**, 566-571.
- Zavalloni, Marisa & Cook, S. W. 1965. Influence of judges' attitudes on ratings of the favorableness of Statements about a social group. *Journal of Personality and Social Psychology*, **1**, 43-54.
- Zax, M. & Louiselle, R. M. 1960. Stimulus values of Rorschach inkblots as measured by the semantic differential. *Journal of Clinical Psychology*, **16**, 160-163.
- Zax, M., Gardiner, D. H. & Lowy, D. G. 1964. Extreme response tendency as a function of emotional adjustment. *Journal of Abnormal and Social Psychology*, **69**, 654-657.
- Zippel, B. 1967. Semantic differential measures of meaningfulness and agreement of meaning. *Journal of Verbal Learning and Verbal Behavior*, **6**, 222-225.
- Young, D. D. 1974. The semantic differential: Application as an affective measure. *Journal of Experimental Education*, **42**, 86-91.

5. Kapitel

Fragebogenkonstruktion

Ulrich Tränkle

1. Einführung

Die Konstruktion eines Fragebogens wird entscheidend davon bestimmt, welche Arten von Informationen (Inhalten) erfaßt, welche Arten von Aussagen gemacht, wofür und auf welcher Grundlage Validität für die Antworten in Anspruch genommen, in welcher Kommunikations-(Befragungs-)form der Fragebogen verwendet werden soll und welche Determinanten des Antwortverhaltens in der Befragungssituation mutmaßlich wirksam sind. Es ist deshalb unerlässlich, in einem einleitenden Kapitel relativ ausführlich auf diese grundlegenden Sachverhalte einzugehen, bevor die mehr technischen Aspekte der Fragenkonstruktion und des Fragebogaufbaus in der Differenziertheit behandelt werden können, die einem häufig verwendeten Instrument wissenschaftlicher Datenerhebung angemessen ist.

1.1 Versuch einer Systematik von Fragebogen

1.1.1 Einteilungsgesichtspunkte für Fragebogen

Fragebogen lassen sich nach einer Vielzahl von Gesichtspunkten einteilen bzw. charakterisieren. Die wichtigsten sind im folgenden zusammengestellt.

- a. Nach dem Grad der Standardisierung lassen sich unterscheiden
 - nicht oder schwach standardisierte Fragebogen
Sie enthalten die Befragungsthemen(-inhalte), aber weder eine genaue Festlegung der Fragen, der Fragenreihenfolge, noch die Antwortmöglichkeiten. Im allgemeinen spricht man hier weniger von Fragebogen als von Interviewerleitfaden, wie sie z.B. in freien Explorationen Verwendung finden.
 - Teilstandardisierte Fragebogen
Sie enthalten in der Regel eine Festlegung von Fragenformulierungen

und Fragenreihenfolge, nicht aber eine Formulierung von Antwortmöglichkeiten.

- Vollstandardisierte Fragebogen

Bei diesen Fragebogen sind die Fragenformulierungen, die Fragenreihenfolge und die Antwortformulierungen festgelegt.

Mischformen zwischen diesen Typen sind möglich und gebräuchlich.

b. Nach der Kommunikationsform werden meist Fragebogen für schriftliche und solche für mündliche Befragung (Atteslander 1971) oder solche für persönliches Interview und schriftliche Befragung (Scheuch 1973) unterschieden. Eine erschöpfende Klassifizierung nach Kommunikationsformen müßte aber mindestens unterscheiden (Tränkle 1974)

- Fragebogen zur Bearbeitung in Anwesenheit eines Interviewers,
 - im Einzelversuch (ein Befragter),
 - mit mündlicher Vorgabe der Fragen,
 - mit mündlicher Beantwortung (persönliches mündliches Interview),
 - mit schriftlicher Beantwortung,
 - mit schriftlicher Vorgabe der Fragen und schriftlicher Beantwortung (persönliches schriftliches Interview),
 - im Gruppenversuch (mehrere Probanden gleichzeitig),
 - mit mündlicher Vorgabe der Fragen und schriftlicher Beantwortung,
 - mit schriftlicher Vorgabe der Fragen und schriftlicher Beantwortung,
- Fragebogen zur Bearbeitung in Abwesenheit eines Interviewers,
 - im Einzelversuch,
 - mit mündlicher Vorgabe der Fragen und mündlicher Beantwortung (z.B. Telefoninterview),
 - im Einzel- oder Gruppenversuch (undefiniert),
 - mit schriftlicher Vorgabe der Fragen und schriftlicher Beantwortung (postalische Befragung).

Wiederum sind Mischformen gebräuchlich, außerdem enthält das Klassifikationsschema nicht alle möglichen Kombinationen der beteiligten Gesichtspunkte, sondern nur diejenigen, für die dem Autor Anwendungen bekannt sind.

c. Nach dem angestrebten Gültigkeitsbereich der Aussagen lassen sich unterscheiden

- individual-diagnostische Fragebogen, die Aussagen über Individuen zum Ziel haben, und
- Fragebogen, die Aussagen über Gruppen (Populationen) anstreben und bei denen das Antwortverhalten des einzelnen Individuums als solches nicht interessiert. Solche Fragebogen werden im folgenden der Kürze

halber als ‚demoskopische‘ oder sozialwissenschaftliche Fragebogen angesprochen.

Hier sind insofern Übergänge möglich, als Aussagen über Populationen auch ausgehend von solchen über Individuen gemacht werden können, in manchen Fällen sogar müssen (vgl. Feger 1974).

- d. Nach dem Inhalt der angestrebten Aussagen kann man unterscheiden
- fakten-, wissens- oder kenntnisorientierte Fragebogen mit individual-diagnostischer (wie bestimmte Intelligenztests) oder demoskopischer Intention,
 - meinungs- bzw. einstellungsorientierte Fragebogen, ebenfalls entweder mit individual-diagnostischer oder demoskopischer Zielsetzung, und
 - persönlichkeitsorientierte diagnostische Fragebogen; hierunter fallen Z.B.
 - Problemfragebogen (adjustment inventories), bei denen es darum geht, das ‚Problemniveau‘ (die Auffälligkeit) einer Person festzustellen,
 - eigenschafts-(trait-)orientierte Fragebogen, die die Messung des Ausprägungsgrades bestimmter Persönlichkeitsmerkmale zum Ziel haben, und
 - Interessenfragebogen, die einen eng umschriebenen inhaltlichen Aspekt der Persönlichkeit, nämlich Vorlieben bzw. Bevorzugungen von Tätigkeiten, Situationen, Berufen zu erfassen trachten (Mitten-ecker 1971).

Auch im Hinblick auf den Inhalt des Fragebogens sind natürlich Mischformen möglich.

- e. Nach dem Grundkonzept der Fragebogenkonstruktion (dem der Konstruktion zugrundeliegenden Validitätskonzept) kann man unterscheiden
- rationale (Anastasi 1968, Edwards 1970), inhaltsorientierte, ‚sample-approach‘-Fragebogen (Cronbach 1970),
 - empirische (Anastasi 1968), statistische, ‚sign-approach‘-Fragebogen (Cronbach 1970),
 - konstrukt-valide, theoriegeleitete (Cronbach 1970) Fragebogen.

Auf diese Grundkonzeptionen von Fragebogen wird im folgenden etwas ausführlicher eingegangen.

1.1.2 Grundkonzeptionen von Fragebogen

Rationale Fragebogenkonstruktionen bestehen in einer Zusammenstellung von Items nach inhaltlichen Gesichtspunkten. Sie wird als repräsentative Stichprobe (Sample) aus einem Universum interessierender Inhalte angesehen. Der Befragte soll die Items verstehen und sie wahrheitsgemäß beantworten. Dem-

entsprechend werden die Antworten ihrer inhaltlichen Bedeutung nach interpretiert und gegebenenfalls zu einem Gesamtscore zusammengefaßt. Dabei werden unter Umständen auch die Interkorrelationen der Items in Betracht gezogen.

Dieser Konstruktionsansatz liegt üblicherweise (vgl. 1.1.3) demoskopischen Fragebogen zugrunde, war aber auch Ausgangspunkt der ersten diagnostischen Fragebogen von Woodworth und Mitarbeitern (Cronbach 1970, Mitenecker 1971).

Innerhalb der rationalen Fragebogenkonstruktion unterscheiden manche Autoren (vgl. Burisch 1976) ein intuitiv-rationales von einem deduktiven Vorgehen bei der Formulierung der Items (Ableitung der Items aus einem - möglicherweise spekulativen - Persönlichkeitsmodell). Sie heben davon einen sogenannten internalen (z.B. Hornick et al. 1977) bzw. induktiven (z.B. Burisch 1976) Ansatz der Konstruktion von Fragebogen auf der Basis der Ergebnisse von Faktorenanalysen der Iteminterkorrelationen ab. Dabei handelt es sich u.E. jedoch nicht um eine eigenständige Fragebogenkonzeption, sondern um eine Technik in der Regel inhaltlicher Validierung. Bekanntlich hängt das Ergebnis einer Faktorenanalyse entscheidend von den einbezogenen Variablen (Items) ab (für den Fall von Persönlichkeitsfragebogen und von zugehörigen Persönlichkeitstheorien hat Coan 1964 dies auch empirisch demonstriert, vgl. auch Scheier & Cattell 1965). über die Einbeziehung eines Items in die Faktorenanalyse wird aber nach inhaltlichen Gesichtspunkten oder nach seiner kriteriumsbezogenen Validität entschieden. Im letztgenannten Fall ist Ziel der Faktorenanalyse ebenfalls die nähere Untersuchung des Fragebogeninhalts.

Prinzipielles Problem des rationalen Konstruktionsansatzes ist, daß er mit durchschaubaren Items arbeitet und arbeiten muß und daß dadurch die Antworten leicht verfälschbar sind (vgl. 1.2.1).

Dies war historisch gesehen auch der Anlaß für die Entwicklung des *empirischen* (bzw. externalen, vgl. Hornick et al. 1977) Konstruktionsansatzes. Wird er in reiner Form verwirklicht, so orientiert sich die Zusammenstellung der Items ausschließlich an ihren Korrelationen zu externen Kriterien. Die Antworten werden als verbales Verhalten betrachtet, das als Zeichen (sign) bzw. Indikator für einen Sachverhalt anzusehen ist, d.h. die Bedeutung einer Antwort ergibt sich allein aus der Korrelation zu Außenkriterien. Auf dieser Grundlage wurden z.B. der MMPI und der Interessenfragebogen von Strong entwickelt. Da die Iteminhalte prinzipiell unerheblich sind, ist es hier möglich, nicht durchschaubare Items zu verwenden und dadurch die Möglichkeit von Verfälschungen erheblich zu reduzieren. Inhaltlich valide Items sind aus diesem Grund für einen streng empirischen Fragebogen geradezu unerwünscht. Die Zusammenfassung von Antworten zu Gesamtscores erfolgt gegebenenfalls nach Maßgabe gemeinsamer Korrelationen mit Außenkriterien. Hauptpro-

bleme dieses Ansatzes sind einerseits die fehlende oder geringe face-validity der Items, wodurch die Motivation der Probanden beeinträchtigt werden kann, andererseits die Risiken, die in einer inhaltsblinden Suche nach empirisch validen Items stecken: Recht häufig wird man dadurch Items mit überhöhten Validitäten in den Fragebogenentwurf aufnehmen und bei Kreuzvalidierungen erhebliche Validitätsrückgänge feststellen. Gelegentlich wird sogar die Auffassung vertreten, „. . . daß substantielle und stabile Zusammenhänge immer nur für Merkmale mit plausibler inhaltlicher Beziehung zu finden . . .“ seien (Burisch 1976, 28). Jedenfalls sind von empirischen Konstruktionen auch nur mittlere Validitäten erreicht worden (Cronbach 1970). Versuche, zur Verbesserung der Akzeptabilität den Items eine von der wirklichen Validität verschiedene face-validity zu verleihen, sind ihrerseits problematisch, da sie evtl. Verfälschungen nach Maßgabe der face-validity begünstigen.

Für praktische diagnostische Anwendungen ist es in jedem Falle unerlässlich, daß ein Fragebogen auch empirisch validiert und nicht, wie bei ausschließlich rationalen Konstruktionen, nur eine „. . . spekulativ halbwegs sinnvoll erscheinende Zusammenstellung . . .“ von Items (Wottawa 1980, 211) ist. Umgekehrt kann kaum ein Testanwender der Versuchung widerstehen, entgegen den Intentionen des Konstrukteurs einen rein empirischen Fragebogen auch inhaltlich zu interpretieren (Cronbach 1970), so daß Überlegungen zur inhaltlichen Validität erforderlich werden. In der Praxis reduziert sich der Unterschied der Grundkonzeptionen häufig auf einen solchen beim ersten Schritt der Item-Selektion (Trennschärfe vs. Kriteriumskorrelation). Darüber hinaus ließ sich auch bei strenger Verwirklichung eine generelle Überlegenheit des einen oder anderen Ansatzes nicht nachweisen (Hase & Goldberg 1967, Burisch 1976, Hornick et al. 1977).

Für *theoriegeleitete* bzw. *konstrukt-valide* Fragebogen (Cronbach & Meehl 1955) ist zum Zwecke der Validierung der Nachweis zu führen, daß es sich bei dem, was sie messen, um ein im Rahmen einer Theorie definiertes Konstrukt handelt. Dieser Nachweis erfolgt vor allem durch Ableitung von Beziehungen zu weiteren Konstrukten aus der Theorie und Überprüfung der Verträglichkeit dessen, was der Fragebogen erfaßt, mit diesen Vorhersagen. Ansätze zu derartigen Fragebogen sind zunächst vor allem von Eysenck (vgl. Eysenck 1953) vorgelegt worden, dessen Intention stets die Messung von Konstrukten war, die in eine umfassende Theorie kortikaler Prozesse eingebaut sind. Demgegenüber kann man allein aufgrund von Versuchen einer Abstraktion von traits aus Fragebogendaten z.B. mittels Faktorenanalyse (vgl. vor allem Guilford 1965) wohl noch nicht von theoriegeleiteter Fragebogenkonstruktion oder von einer Konstruktvalidierung von Fragebogen sprechen (Cronbach 1970). Die faktorielle Reinheit von Fragebogenitems hat - entgegen Holm (1974 a, b) - nichts mit ihrer Theorieorientiertheit zu tun. Außerdem gibt es kaum sachliche (allenfalls technische) Gründe, sie zu fordern (Cattell 1974),

zumal sie ja nicht an sich, sondern immer nur innerhalb einer gegebenen Variablen-(Item-)stichprobe existiert. Von erheblichem Nutzen kann die Faktorenanalyse sein, wenn ein inhaltlich (rational) konstruierter Fragebogen z.B. durch Einbeziehen von Markierungsvariablen daraufhin überprüft werden soll, ob die angestrebten Inhalte tatsächlich enthalten sind (so schon Eysenck 1953). Außerdem läßt sich mit ihrer Hilfe bei einem empirischen Fragebogen, besonders wenn die Kriterien mit in die Analyse einbezogen werden, die zunächst nur statistische Beziehung zwischen Antwort und Kriterien auch in ihrer inhaltlichen Bedeutung erhellen, d.h. ein zunächst nur empirisches in ein auch inhaltliches Validitätskonzept überführen.

1.1.3 Hauptanwendungsgebiete für Fragebogen

Diagnostische Fragebogen sind Tests, insofern orientiert sich ihre psychometrische Konstruktion an einem Test- bzw. Meßmodell. Bezüglich der damit zusammenhängenden Fragen muß auf die einschlägige Literatur, z.B. Gulliksen (1950), Magnussen (1966), Lienert (1969), Fischer (1974), verwiesen werden. Faßt man einen Fragebogen ganz allgemein als eine Zusammenstellung von Fragen auf, so enthalten die meisten Tests auch Fragebogen. Für Fragen in Leistungstests ist charakteristisch, daß es für sie eine objektiv richtige Antwort gibt, so daß besondere Überlegungen zum Problem des Ratens erforderlich werden (vgl. Wottawa 1980). Diagnostische Fragebogen im engeren Sinne sind z.B. Persönlichkeits- und Interessenfragebogen, für ihre Items gibt es höchstens subjektiv richtige Antworten. Die folgenden Ausführungen beschränken sich in der Regel auf Fragebogen dieses Typs. Fast immer sind Persönlichkeits- und Interessenfragebogen vollstandardisierte Verfahren, mit schriftlicher Vorgabe von Fragen und schriftlicher Beantwortung in Anwesenheit eines Untersuchers, die Durchführung erfolgt teils im Einzelversuch, teils in Gruppen.

Demoskopische Fragebogen können alle möglichen Standardisierungsgrade aufweisen. Die Befragungen werden meist als persönliche mündliche Interviews im Einzelversuch oder ‚postalische‘ Befragungen, seltener als persönliche schriftliche Interviews durchgeführt. Die Fragebogen haben Fakten, Wissen, Kenntnisse, Meinungen oder Einstellungen zum Inhalt und sind vom Validitätskonzept her fast ausschließlich rationale bzw. inhaltsorientierte Konstruktionen. Überlegungen zu empirischen (kriteriumsorientierten) Validierungen werden vor allem im Zusammenhang mit Meinungs- und Einstellungsfragen und ihrer Indikatorfunktion angestellt (Friedrich 1971), soweit Fakten- und Wissensbereiche thematisch sind, interessieren eher Verfälschungen bzw. Fehler (Lansing et al. 1961) und ihre Hintergründe, bzw. Fragen der Reproduktion oder des Wiedererkennens von Gedächtnisinhalten (Cannell et al. 1977).

In gewisser Hinsicht eine Zwischenstellung zwischen diagnostischen (Persönlichkeits- und Interessen-) und demoskopischen Fragebogen besitzen die sogenannten Einstellungs-(Attitüden-)skalen, die eingehend etwa bei Edwards (1957), Scheuch (1962), Süllwold (1969), Scheuch & Zehnpfennig (1974) behandelt werden. Wie demoskopische Fragebogen werden sie fast ausnahmslos mit dem Endziel von Aussagen über Gruppen und nicht zu individual-diagnostischen Zwecken eingesetzt, doch sind sie - von Eigentümlichkeiten der Itemselektion abgesehen - formal mit diagnostischen Fragebogen identisch. Die Konstruktion von Einstellungsskalen geht stets von Iteminhalten aus, erfordert aber mindestens bei Skalen des Thurstone-Typs auch eine empirische Validierung. Während für die Lickert-Skalen die Trennschärfe unter Zugrundelegung von Gesamtscore-Extremgruppen (also ein inhaltliches Konzept) das Selektionskriterium für Items darstellt, werden bei den Thurstone-Skalen die Items unter Verwendung einer ‚Eichstichprobe‘ hinsichtlich der Extremität der durch sie zum Ausdruck gebrachten Einstellung skaliert. Skalenwert des Individuums ist im ersten genannten Falle die Summe der graduell abgestuften Zustimmungen zu den Items, im letztgenannten Falle (Modifikationen dieser Vorgehensweise einmal unberücksichtigt gelassen) der Skalenwert des Items, das der Proband am ehesten für zutreffend hält. Abgesehen von Einwänden, die sich auf die ungeprüft unterstellte Eindimensionalität der gemessenen Sachverhalte beziehen und denen Guttman mit der Skalogrammanalyse zu begegnen versuchte (vgl. Edwards 1957), wäre bei der Verwendung von Zustimmungsgraden in Lickert-Skalen die Dimensionalität der Antworten, z.B. durch Zugrundelegung eines mehrkategoriiellen probabilistischen Meßmodells zu berücksichtigen (Wottawa 1980); jedenfalls läßt sich die mehr oder weniger willkürliche Verwendung von (gleichabständigen) Gewichten für die Zustimmungsgrade bei der Bildung eines Gesamtscores kaum rechtfertigen. Für Thurstone-Skalen ist ungeklärt, wie sich Personen und an einer Eichstichprobe skalierte Items in einem gemeinsamen psychologischen Raum darstellen lassen könnten. Eine kritische Auseinandersetzung mit den Ansätzen von Thurstone, Lickert und Guttman sowie alternative Vorgehensweisen finden sich z.B. bei Feger (1974) und Lantermann & Gehlen (1977).

Trotz der vorstehend beschriebenen Akzentuierungen gibt es hinsichtlich zahlreicher Probleme keine prinzipiellen Unterschiede zwischen diagnostischen Fragebogen, Einstellungsskalen und demoskopischen Fragebogen, so daß die nachfolgenden Ausführungen sich nur ausnahmsweise explizit auf bestimmte Anwendungssituationen beziehen.

1.2 Ansätze zu einer Theorie des Beantwortungsprozesses

1.2.1 *Determinanten des Antwortverhaltens*

Als erster Schritt auf dem Wege zu einer Theorie des Beantwortungsprozesses bietet sich die Analyse dessen an, was bei der Entstehung einer Antwort in der Vp vor sich geht. Entsprechende Untersuchungen sind für den Fall von Persönlichkeitsfragebogenitems mehrfach durchgeführt worden (vgl. etwa die Nachweise bei Cronbach 1970, Schneider-Düker & Schneider 1977, Kalinowsky-Czech 1979), ihre Ergebnisse dürften sich prinzipiell aber auch auf demoskopische Fragen übertragen lassen. Turner & Fiske (1968) und in Fortführung dieses Ansatzes Kuncel (1973, vgl. auch Fiske 1978) und ebenso Nowakowska (1971) verwendeten einen ‚Mets-Fragebogen‘ zur nachträglichen Erfassung der Beantwortungsprozesse, Schneider-Düker & Schneider (1977) und Kalinowsky-Czech (1979) bedienten sich der Methode des ‚Lauten Denkens‘, die letztgenannte Autorin ließ ihre Vpn außerdem frei zu den Items assoziieren. Rogers (1974 a, b) variierte bestehende Charakteristika der Items experimentell und zog ausgehend von den Veränderungen der Beantwortungszeiten (Reaktionszeiten) Schlüsse auf den Beantwortungsprozeß. Cliff et al. (1973) und Cliff (1977) versuchten, Beziehungen zwischen der Beantwortung von Items und ihrer Bedeutung (ausgehend von der MDS ihrer Ähnlichkeitsstruktur) herzustellen. übereinstimmend zeigten diese Untersuchungen, daß einerseits das Verständnis ein- und derselben Frage von Vp zu Vp beträchtlich variiert und andererseits die Beantwortungsprozesse ein- und derselben Vp itemspezifisch recht unterschiedlich ablaufen (vgl. auch Crutchfield & Gordon 1947). Turner & Fiske (1968) etwa klassifizierten ausgehend von MMPI-Items nur etwa 50% der von Vpn beschriebenen Beantwortungsprozesse als adäquat im Sinne der Intention des Fragebogens bzw. seiner Autoren. Dazu dürfte u.a. beitragen, daß die Vorstellungen, die Vpn mit den in Fragebogenitems häufig anzutreffenden unbestimmten Zahlen- und Häufigkeitsangaben (‚einige‘, ‚manchmal‘) verbinden, sehr unterschiedlich sind (Simpson 1944, Strahan & Gerbasi 1973, Schriesheim & Schriesheim 1974, Rohrmann 1978, Bradburn & Sudman 1979).

Weitere Erkenntnisse betreffend die Determination des Antwortverhaltens kommen sodann von experimentellen Untersuchungen zur Verfälschbarkeit (faking) von Antworten auf Fragebogenitems, z.B. in vorgestellten Situationen, und von Untersuchungen zur Wirksamkeit von Antworttendenzen bei der Bearbeitung von Fragebogen in Ernstsituationen. Auf die kaum noch überschaubare Fülle der dazu vorliegenden empirischen Befunde kann an dieser Stelle nicht näher eingegangen werden, zusammenfassend orientieren darüber z.B. Block (1965), Berg (1967), Anastasi (1968), Cronbach (1970), Edwards (1970). Im deutschen Sprachraum sind insbesondere Untersuchungen von Cohen & Carl (1964), Carl (1968), Fürntratt (1969), Tholey (1976), Ham-

pel & Klinkhammer (1978), Jannssen (1978), Häcker et al. (1979) und mehrere Arbeiten von Hoeth und Mitarbeitern (zusammenfassend Hoeth 1980) zu nennen.

Mit Möglichkeiten der Vermeidung bzw. Erfassung und Elimination der Einflüsse von Antworttendenzen (Reaktionseinstellungen, response sets) und den Schwierigkeiten ihrer Realisierung bei der Fragebogenkonstruktion setzen sich u.a. Ehlers (1973), Janke (1973) und Keil (1973) auseinander.

Als *Antworttendenzen* könnte man allgemein diejenigen systematischen Anteile im Antwortverhalten der Vpn bezeichnen, die nicht auf den jeweiligen (subjektiv) wahren Sachverhalt, sondern auf die Form der Frage bzw. der Befragung insgesamt zurückzuführen sind. Untersucht wurden derartige Antworttendenzen vor allem im Zusammenhang mit diagnostischen Fragebogen, doch lassen sie sich nach Hoeth (1980) auch für demoskopische (sozialwissenschaftliche) Fragebogen leicht aufzeigen.

Bei den *frageninhaltsorientierten* Antworttendenzen wie Simulation, Dissimulation und der besonders intensiv untersuchten Tendenz zu sozial erwünschten Antworten (SD-Tendenz) werden die Antworten auf Fragen im Hinblick auf ganz bestimmte Zwecke (z.B. einen ‚guten Eindruck‘ zu machen) verfälscht.

Demgegenüber erfolgt bei den *antwortinhaltsorientierten* Antworttendenzen (den response sets im engeren Sinne) eine Bevorzugung von Antworten ganz bestimmten Inhalts ohne Rücksicht auf die Inhalte der Fragen. Am meisten Aufmerksamkeit auch im Hinblick auf Beziehungen zu bestimmten Persönlichkeitsmerkmalen hat dabei die Bejahungs- oder Zustimmungstendenz (acquiescence) gefunden, außerdem wurden Verneinungstendenzen, Mittentendenzen, Extremtendenzen und Variationstendenzen aufgezeigt.

Schließlich gibt es *nicht-inhaltsorientierte* Antworttendenzen, dazu gehören Positionseffekt und formale Antwortstereotypien (Antwortmuster).

Antworttendenzen und Unterschiede im Verständnis von Fragebogenitems stellen im Rahmen einer rationalen (inhaltsorientierten) Fragebogenkonstruktion (sample-approach) sicherlich ein schwerwiegendes Problem dar (Eysenck 1953), das durch geeignete Formulierung und Zusammenstellung von Items bzw. durch Verwendung spezieller Fragen oder Fragentypen (forced-choice-items) bestenfalls gemildert, nicht jedoch überwunden werden kann (Ehlers 1973, Keil 1973).

Ihrer Natur nach sind Verständnisunterschiede und Antworttendenzen nicht als intraindividuell unkorrelierte Zufallsfehler mit einem Erwartungswert von Null anzusehen, so daß entsprechend dem Reliabilitätskonzept der klassischen Testtheorie erwartet werden dürfte, daß sie sich mit steigender Zahl homoge-

ner Items zur Erfassung des jeweiligen Sachverhaltes tendenziell aufheben. Ein probabilistisches Meßmodell bietet zwar prinzipiell den Vorteil, daß sich hier im individual-diagnostischen Anwendungsfall von den Eigentümlichkeiten der verwendeten Items (Item-Parametern) befreite spezifisch objektive Personen-kennwerte (Personen-Parameter) bestimmen lassen, im Falle von gruppenskriptiven Zielsetzungen Kennwerte @em-Parameter), die von den verwendeten Vpn (Personen-Parametern) bereinigt sind (Sixtl 1972, Andersen 1973, Fischer 1974), doch würden Antworttendenzen und Verständnisunterschiede der Items bei verschiedenen Vpn gravierend gegen die Modellannahme der Unabhängigkeit der Item- von den Personen-Parametern verstoßen. Innerhalb eines inhaltsorientierten (rationalen) Ansatzes zeichnen sich also keine Möglichkeiten ab, das Problem der Antworttendenzen und Verständnisunterschiede für Items zu lösen (im Zusammenhang mit Überlegungen zur Fragenformulierung wird auf diesen Punkt noch einmal zurückzukommen sein, vgl. 3.).

Im Rahmen des rein empirischen Validitätskonzeptes für Fragebogen (sign-approach) stellen die Unterschiede im Verständnis der Fragen kein prinzipielles Problem dar. Vielmehr ist durchaus denkbar, daß sie eine wesentliche Grundlage der empirischen Validität sind und daß diese sinken würde, wollte man die Verständnisunterschiede zwischen Vpn reduzieren (Cronbach 1970, Mittenecker 1971).

So berichten Strahan & Gerbasi (1973) tatsächlich deutliche Korrelationen zwischen Interpretationen von Items und Persönlichkeitsdimensionen (die sie freilich anders interpretieren). Die Minimalbedingung für die Brauchbarkeit von Antworten in Fragebogen ist nur, daß sie „... irgendwie psychologisch bedeutsame Reaktionen . . .“ sein müssen (Mittenecker 1971, 480).

Solche Reaktionen liegen nur dann nicht vor, wenn die Vp auf Items eines Fragebogens ohne Bezug zum Inhalt von Frage- oder Antwortmöglichkeiten reagiert, d.h. z.B. zufällig oder willkürlich antwortet. Verfälschungstendenzen (wie soziale Erwünschtheit) bzw. responsesets (wie acquiescence) können dagegen als solche psychologisch bedeutungsvoll sein und schließen die Verwendbarkeit von Reaktionen nicht a priori aus (Cattell 1974). Unter dem Begriff ‚response style‘ hat man insbesondere die Zustimmungstendenz (acquiescence) selbst zum Persönlichkeitsmerkmal erhoben, also das, was ursprünglich eine Verfälschung der Antworten zu sein schien, zu einer inhaltlich interessierenden Fragebogenvariablen gemacht (vgl. als kritische Übersicht Rorer 1965).

1.2.2 Antwortgenese

Die prinzipiellen Möglichkeiten für das Zustandekommen der Antwort auf ein Fragebogenitem sind in Abb. 1 in Form eines Flußdiagramms dargestellt, in

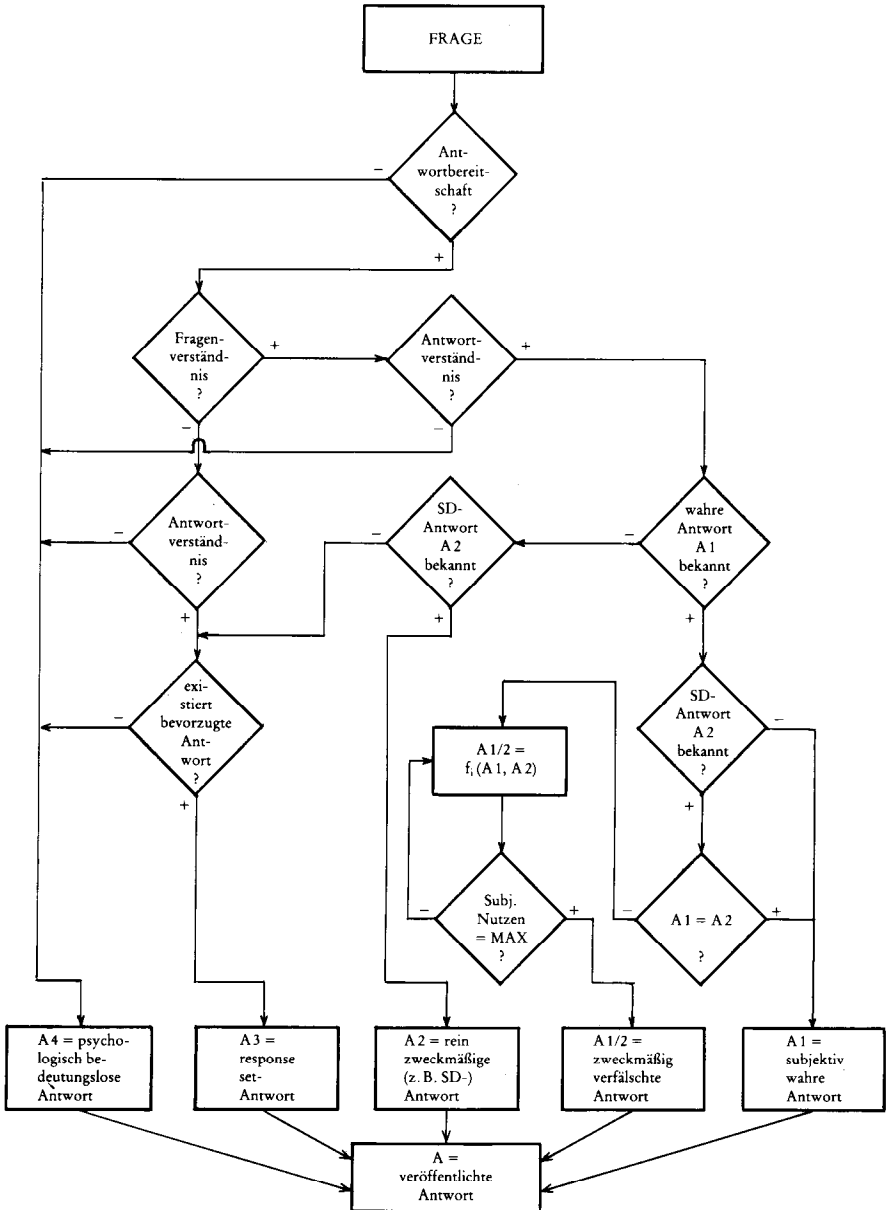


Abb. 1: Modell der Antwortgenese

das Vorstellungen von Getzels (1954), Damarin (1970), Nowakowska (1971), Schneider (1972) eingegangen sind. Dabei handelt es sich um ein Modell der Antwortgenese, nicht etwa um eine phänomenologische Beschreibung, d.h. es wird nicht angenommen, daß die Antwortgenese sich im Bewußtsein der Vp so darstellt, wie das Modell sie abbildet.

Eine Antwort in dem hier zugrundegelegten Sinne ist auch die Nichtbeantwortung (das Ausbleiben einer inhaltlichen Antwort) bzw. die Enthaltung (Wahl einer Neutralkategorie).

Bei der Entstehung einer Antwort gibt es zunächst die Möglichkeit, daß die Vp, z.B. weil sie nicht zu angemessener Mitarbeit bereit ist, vom Inhalt des Items bzw. von den Inhalten der Antwortmöglichkeiten gänzlich unbeeinflußt reagiert. Dies führt zu einer mindestens im Zusammenhang mit dem Item psychologisch nicht bedeutungsvollen, ‚zufälligen‘ oder willkürlich systematischen Antwort (A4). Derartige Antworten können die Grundlage sogenannter Positioneffekte sein. Natürlich besteht die Möglichkeit, aus dem Beantwortungsprozeß ‚auszusteigen‘ und nach A4 zu verzweigen, auf allen nachgeordneten Stufen. Im Interesse der Übersichtlichkeit haben wir dies im Flußdiagramm nicht eigens vorgesehen (wollte man es tun, könnte man nach jedem Schritt eine weitere Abfrage ‚Motivation noch ausreichend‘ o.ä. einbauen).

Sodann kann eine Antwort zustande kommen durch die Verarbeitung von Frage und explizit oder implizit vorgegebenen Antwortmöglichkeiten unter Heranziehung von Gedächtnisinhalten (Wissen, Normen). Dies setzt voraus, daß ein subjektives (d.h. von der Vp als solches erlebtes) Verständnis von Frage und/oder Antwortmöglichkeiten erzielt worden ist. Sieht man von A4, der psychologisch nicht bedeutungsvollen Antwort, ab, so kann die Vp ihre Antwort nach dem Kriterium der subjektiven Richtigkeit (Wahrheit) als richtige Antwort (A1), in Verfolgung eines bestimmten Zweckes z.B. entsprechend der (subjektiven) sozialen Erwünschtheit (A2) oder nach Maßgabe einer bei ihr vorherrschenden Reaktionstendenz (response set), also ohne Rücksicht auf den Fragen-, aber mit Bezug auf den Antwortinhalt auswählen (A3). Alle diese Möglichkeiten entsprechen der Minimalbedingung der psychologischen Bedeutsamkeit (Mittenecker 1971) für Antworten.

Ist eine Frage für eine Vp bedeutsam, d.h. trifft sie auf ihre eigene Situation zu und verbindet sie mit der Frage kognitive Inhalte oder positive bzw. negative Assoziationen, so ist sie auch in der Lage, zu bestimmen, welche Antwort die subjektiv richtige (A1) ist (Nowakowska 1971). Diese Antwort wird sie unmittelbar als endgültige Antwort (A) verlautbaren, wenn sie keine Kriterien (spezifische Erfahrung mit Antwortkonsequenzen, verinnerlichte gesellschaftliche Normen) für die Zweckmäßigkeit (z.B. für die soziale Erwünschtheit) einer Antwort besitzt. Besitzt sie solche Kriterien, so wird sie zusätzlich eine ‚zweckmäßige‘ (d.h. meist sozial erwünschte) Antwort (A2) entwerfen, die

mit der subjektiv richtigen Antwort übereinstimmen ($A1 = A2$) oder von dieser abweichen kann. Daneben gibt es auch die Möglichkeit, daß die Vp eine subjektiv richtige Antwort (z.B. mangels Betroffensein) nicht kennt, jedoch weiß, welche Antwort zweckmäßig, z.B. sozial erwünscht ist. In diesem Fall wird sie eine sozial erwünschte Antwort ($A2$) wählen (die natürlich auch in einer Nichtbeantwortung oder Enthaltung bestehen kann). Einerseits unterscheiden sich Probanden in der Neigung, relativ unabhängig von der konkreten Frage sozial erwünscht zu antworten, andererseits unterscheiden sich aber auch Fragen in der Neigung, relativ unabhängig von den Probanden sozial erwünscht beantwortet zu werden. Holm (1974 b) differenziert dementsprechend im Rahmen seiner faktorenanalytischen Theorie der Frage bzw. Fragenbatterie zwischen einer ‚allgemeinen sozialen Erwünschtheit‘ und einer (fragen-) ‚spezifischen sozialen Erwünschtheit‘. Letztere wird häufig auch als Suggestivwirkung einer Frage bezeichnet.

Stehen sich zwei verschiedene Antwortmöglichkeiten, eine subjektiv richtige ($A1$) und eine zweckmäßige, z.B. sozial erwünschte ($A2$) gegenüber, so trifft die Vp nach den Ergebnissen von Nowakowska (1971), die sie mittels Faktorenanalyse von Zusammenhängen in den Beschreibungen der Beantwortungsprozesse bei 28 Items aus dem 16 PF von Cattell gewann, die Entscheidung in Abhängigkeit von der subjektiven Nützlichkeit (N) der Antwort. Die subjektive Nutzenfunktion wird dabei durch die erwarteten materiellen und gesellschaftlichen Konsequenzen der Antwort einerseits und die Konsequenzen für das Selbstwertgefühl (z.B. bei Abweichung von der Wahrheit) andererseits bestimmt. Die Optimierung unter dem Kriterium des subjektiven Nutzens kann zur Wahl von $A1$ oder $A2$ führen oder eine kombinierte Antwort ($A1/2$) = $f(A1, A2)$ erzeugen, also eine durch ‚Zweckmäßigkeitsüberlegungen verfälschte richtige bzw. eine in dem Bestreben nach Wahrheit veränderte Zweckantwort. Diese Veränderung der Antwortentwürfe kommt für die Vp natürlich nur in Betracht, wenn $A1$ und $A2$, d.h. die wahre und die zweckmäßige Antwort, divergieren.

Fragen, bei denen die Vp erhebliche gesellschaftliche Konsequenzen im Falle einer sozial nicht erwünschten Beantwortung erwartet, dürften den größten Teil dessen abdecken, was in der Literatur unter den Bezeichnungen ‚schwierige‘, ‚heikle‘, ‚unangenehme‘ Fragen abgehandelt wird. Für solche Fragen gilt als typisch, daß sie relativ hohe Nichtbeantwortungsquoten aufweisen. Vermutlich handelt es sich hier um Fälle, in denen der Konflikt zwischen wahrer und zweckmäßiger (sozial erwünschter) Antwort durch Nichtbeantwortung gelöst wird, d.h. in denen die Nichtbeantwortung den höchsten subjektiven Nutzen verspricht. In diese Richtung deuten die Ergebnisse von Koolwijk (1968, 1969), denenzufolge Fragen nicht an sich unangenehm sind, sondern in Abhängigkeit davon, wie die wahre Antwort der Vp ausfallen müßte, in sehr verschiedenem Ausmaß als unangenehm empfunden werden.

Gelingt es der Vp, z.B. aus Mangel an Aktualität (subjektiver Bedeutsamkeit) der Frage und an Vorstellungen über gesellschaftliche Konsequenzen der Antworten nicht, eine am Frageninhalt orientierte Antwort (A1, A2, A1/2) zu entwickeln, so tritt eine vom Frageninhalt unabhängige und - soweit solche vorhanden sind - durch Antwortbevorzugungstendenzen bestimmte Antwort auf. Derartige response sets kommen demnach also nur ins Spiel, wenn eine Vp weder Kriterien für eine richtige, noch solche für eine zweckmäßige Antwort hat. Holm (1974a) spricht in diesem Falle von Fragen ohne klare Zieldimension. Die Tendenz zu sozial erwünschten Antworten hat demgegenüber andere Qualität. Sie tritt nicht nur auf, wenn die Vp eine wahrheitsgemäße Antwort nicht zur Verfügung hat, sondern konkurriert mit dieser.

Sollte bei Unmöglichkeit frageninhaltsorientierten Antwortens (A1, A2, A1/2) die Vp auch nicht über Antwortbevorzugungstendenzen verfügen, die A3 determinieren könnten, tritt eine willkürliche Antwort (A4) auf, bei der es sich natürlich auch um eine Auslassung handeln kann.

Dem Untersucher steht im Normalfall nur die endgültige Antwort (A) der Vp zur Verfügung. Durch Konstruktion des Fragebogens sollte er soweit als möglich sicherstellen, daß bei zugrundegelegtem inhaltlichen Validitätskonzept (sample approach) $A = A1$, bei empirischem Validitätskonzept (sign approach) im Sinne obiger Minimalbedingungen $A \neq A4$ ist.

Für einen Teil der Fragen in demoskopischen Interviews existieren objektiv richtige Antworten. In diesen Fällen ist es gerechtfertigt, von ‚Beantwortungsfehlern‘ (response-errors) zu sprechen. Unter (im Vergleich zu obigem Modell der Antwortgenese) stärkerer Betonung der Ursachen für die Fehler unterscheiden Lansing et al. (1961)

- motivationsbedingte Beantwortungsfehler, die vorliegen, wenn der Proband nicht motiviert ist, die richtige Antwort zu geben, selbst wenn er das könnte (vgl. Cattell 1974),
- kommunikationsbedingte Beantwortungsfehler, die vorliegen, wenn
 - der Proband nicht versteht, welche Information von ihm erwartet wird, d.h. der Fragesteller sich nicht verständlich gemacht hat,
 - der Untersucher die vom Probanden übermittelte Information nicht versteht, d.h. der Proband sich nicht verständlich gemacht hat,
- Unwissenheitsfehler, die vorliegen, wenn dem Probanden die erbetene Information nicht zur Verfügung steht.

Diesen *Beantwortungsfehlern* (response-errors) sind noch die *Antwortverweigerungsfehler* (errors-of-non-response) an die Seite zu stellen, also Fehler, die die Verallgemeinerungsfähigkeit von Befragungsergebnissen betreffen und die ganz besonders im Zusammenhang mit unpersönlichen (z.B. postalischen) Befragungen ein gravierendes Problem sind. Als dritter Fehlertypus neben

Beantwortungs- und Antwortverweigerungsfehlern sind *Stichprobenfehler* in Rechnung zu stellen (Lansing et al. 1961), die - da es sich nicht um spezifisch mit Befragungen bzw. Fragebogen verbundene Fehler handelt - im Rahmen dieser Abhandlung unberücksichtigt bleiben sollen.

1.2.3 Die Frage als Suchbegriff

Aus der eher makroskopischen Betrachtungsweise des Modells für die Antwortgenese soll ein Aspekt wegen seiner Wichtigkeit herausgegriffen und etwas näher beleuchtet werden: die Auffindung der für die subjektiv richtige Antwort erforderlichen Gedächtnisinhalte. Insbesondere bei einer Ausrichtung von Fragen auf inhaltliche Validität wird man sich nicht mit der Abfrage ‚wahre Antwort bekannt‘ (siehe Abb. 1) zufrieden geben, man wird sich vielmehr überlegen müssen, wie die Vp versucht, die wahre Antwort aufzufinden, und wie der Fragebogenkonstrukteur ihr dabei helfen kann. Cannell et al. (1977) haben im Zusammenhang mit der Diskussion von Faktenfragen, also von Fragen, für die inhaltliche Validität angestrebt wird, deutlich gemacht, daß für solche Fragen einerseits die Speicherung der Information im Gedächtnis der Vp und die zu ihrer Auffindung erforderlichen Suchprozesse, andererseits die Vorstellungen des Fragebogenkonstruktors von den zu erfragenden Fakten näher untersucht und in einer formulierten Frage zur Deckung gebracht werden müssen. Abb. 2 (in Anlehnung an Cannell et al. 1977, 53) gibt die dabei mindestens zu beachtenden Schritte wieder.

Eine Gedächtnisspur nimmt nicht vom wahren Sachverhalt, sondern vom Phänomen, d.h. von dem im Erleben des Probanden realisierten Sachverhalt ihren Ausgang, wohingegen der Untersucher seine Vorstellung des Sachverhaltes zugrundelegt. Da der Untersucher zunächst nicht weiß, wie ein Proband einen bestimmten Sachverhalt erlebt und in welcher Kodierung er ihn abgespeichert hat, ist es - um die Wahrscheinlichkeit der Auffindung zu erhöhen - erforderlich, möglichst viele alternative Vorstellungen von dem Sachverhalt zu entwickeln. Um ein Beispiel von Cannell et al. (1977) zu verwenden: Was sich für den Untersucher als ‚zahnärztliche Behandlung‘ darstellt, kann im Erleben des Befragten in erster Linie eine ‚besonders schmerzhaft Erfahrung‘ oder eine ‚hohe finanzielle Belastung‘ gewesen sein. Fragt man nicht nur nach ‚zahnärztlicher Behandlung‘, sondern verwendet auch die beiden alternativen Vorstellungen vom Sachverhalt, so erhöht man die Wahrscheinlichkeit, daß der Proband die gesuchte Information in seinem Speicher auffindet. Aber auch *ein* bestimmter erlebter Sachverhalt kann in sehr verschiedene *Bezugssysteme* eingebettet sein. Bei der Operationalisierung des Sachverhaltes als Untersuchungsvariable kommt es darauf an, möglichst viele der in Frage kommenden Bezugssysteme zu berücksichtigen. Fragt man nach Krankheiten des Probanden nicht nur im Kontext eines Klassifikationsschemas für Krankheiten, son-

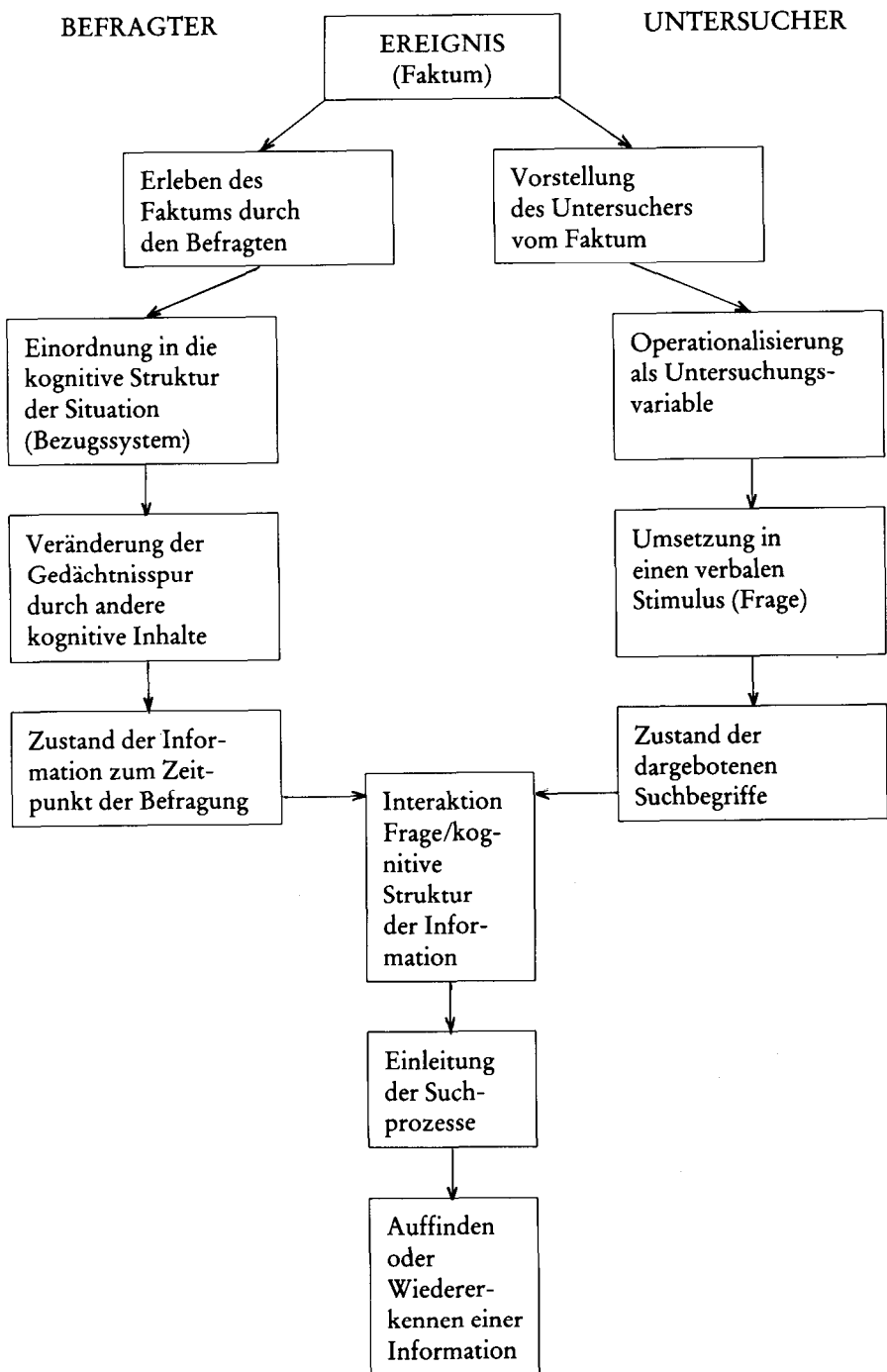


Abb. 2: Interaktion von Frage und kognitiver Struktur der Information

dern z.B. auch auf dem Hintergrund der Bezugssysteme ‚Symptome‘ (Schmerzen), ‚Ereignisse‘ (Krankheitsfälle) und ‚Lebensgewohnheiten‘ (Diät, Medikamenteneinnahme etc.), so lassen sich die Nennungen tatsächlich vorliegender Krankheiten ganz beträchtlich erhöhen (Cannell et al. 1977). Veränderungen der Gedächtnisspur durch andere kognitive Inhalte müssen dabei in Betracht gezogen werden. So kann z.B. ein Krankenhausaufenthalt in der Kindheit durch spätere Krankenhausaufenthalte an subjektiver Bedeutsamkeit verlieren. Bei der Umsetzung der Untersuchungsvariable ‚Krankenhausaufenthalte‘ wäre dies z.B. dadurch zu berücksichtigen, daß man den Probanden fragt, ob ein berichteter Krankenhausaufenthalt tatsächlich der erste Aufenthalt in einem Krankenhaus war.

Die in Abb. 2 aufgelisteten Verarbeitungsprozesse auf seiten des Befragten sollten vom Untersucher stets im Auge behalten werden, wenn er versucht, einen Sachverhalt mit Hilfe von Fragen zu erfassen. Soweit es sich um Wissenstatbestände handelt, muß dabei eine Optimierung der Befragungssituation im Hinblick auf die Wiedergabe (Reproduktion bzw. Wiedererkennen) angestrebt werden, ein Gebiet, das im Vergleich zum optimalen Lernen in der empirischen Forschung bisher recht stiefmütterlich behandelt worden ist (Cannell et al. 1977). Nur dann wird man erreichen können, daß die Beantwortungsfehler, von denen der Untersucher gerne spricht und die er nur in Ausnahmefällen (methodologischen Studien) als solche erkennen kann, nicht in Wahrheit Befragungsfehler sind.

1.3 Einordnung der Fragebogenkonstruktion in die Stadien einer Befragung

Die Konstruktion eines Fragebogens ist eingebettet in den Prozeß der Planung, Vorbereitung, Durchführung und Auswertung einer Befragung oder Testung. Um diese Einbettung deutlich zu machen, führen wir nachstehend die wichtigsten Stadien dieses Prozesses für demoskopische Befragungen auf. Sinngemäß sind sie auch auf diagnostische Untersuchungen übertragbar. Die Fragebogenkonstruktion im engeren Sinne umfaßt dabei Entwurf, Erprobung und Revision eines Fragebogens, nötigenfalls mehrfach wiederholt.

- a. Sichtung und Aufarbeitung der zum Themenbereich vorliegenden theoretischen Ansätze und empirischen Befunde, besonders soweit es sich um Ergebnisse früherer Befragungen handelt.
- b. Formulierung der genauen Fragestellung, soweit es sich um eine Befragung mit hypothesenprüfendem Anspruch handelt auch der Hypothesen, dabei auch explizite Festlegung der Grundgesamtheit, auf die sich die Hypothesen beziehen.
- c. Differenzierung der Fragestellung in einzelne, als Untersuchungsvariablen

geeignete Aspekte (Definition der abhängigen Variablen, der ‚Programmfragen‘ i. S.v. Noelle 1963).

- d. Festlegung der interessierenden und zu erfassenden Kovarianten.
- e. Erweiterung der Informationsbasis z.B. durch Expertenbefragungen, freie Explorationen mit Betroffenen, Gruppendiskussionen, aber auch durch Analyse von Medien u.ä. Dieser Schritt kann evtl. auch schon an früherer Stelle erfolgen.
- f. Soweit erforderlich Revision der in b.-d. getroffenen Festlegungen aufgrund der Erkenntnis in e.
- g. Festlegung der genauen Untersuchungsmethode, vor allem Entscheidung zwischen mündlicher, persönlich-schriftlicher oder unpersönlich-schriftlicher Befragung auf dem Hintergrund der inhaltlichen Festlegungen in den Schritten a.-f.
- h. Erstellung eines Fragebogenentwurfs durch
 - Operationalisierung der Untersuchungsvariablen, d.h. ihre Umsetzung in ‚Ermittlungsfragen‘ (Noelle 1963) unter Berücksichtigung möglicher Formulierungseffekte,
 - Festlegung des Befragungsverlaufes durch Definition der Reihenfolge der Fragen unter Berücksichtigung möglicher Reihenfolgeeffekte,
 - Festlegung der Fragebogengestaltung (Layout) unter Berücksichtigung möglicher Einflüsse auf Antwortbereitschaft und Art der Beantwortung.

Bei der Erstellung des Fragebogenentwurfs sind außer den inhaltlichen Gesichtspunkten und den genannten möglichen Einflüssen auf das Antwortverhalten besonders auch die jeweilige Zielpopulation (Grundgesamtheit) und die genaue Befragungsmethode zu bedenken. Außerdem muß schon in diesem Stadium im Hinblick auf Kodierungen die spätere Auswertung genau geplant werden.

- i. Erprobung des Fragebogens (Pretest) an einer im Vergleich zur Hauptuntersuchung kleineren, aber für die Grundgesamtheit ebenfalls repräsentativen Stichprobe. Im Rahmen dieser Erprobung sollten einerseits Interviewer die Vpn bei der Beantwortung der Fragen möglichst systematisch beobachten, um Verständnisschwierigkeiten besonders durch Fragenformulierungen und Fragenreihenfolge (Verzweigungen) aufzudecken. Zum anderen sollten die Antworten der Vpn verwendet werden
 - zur Entdeckung häufig ausgelassener Fragen,
 - zur Ermittlung und Analyse von Antwortverteilungen (mehrgipflige Verteilungen weisen häufig auf Mehrdeutigkeit der Fragen hin),
 - zu Itemanalysen bzw. -Validierungen, je nach verfolgtem Validitätskonzept (dabei können z.B. Konsistenzanalysen, Stabilitätsbestimmungen, Kriteriumskorrelationen der Items und Faktorenanalysen der Iteminterkorrelationen angezeigt sein),
 - zur Entdeckung von Verfälschungsmöglichkeiten (soziale Erwünschtheit, response sets) und Tendenzen zu unsorgfältiger (zufälliger) Be-

antwortung (dabei ist auch eine Bestimmung des akzeptablen Fragebogenumfangs vorzunehmen und zu ermitteln, ob zusätzliche Instruktionen erforderlich sind).

Auf dieser Stufe, evtl. auch schon bei der Erstellung des ersten Fragebogenentwurfes, können evtl. Sprachanalysen der Fragenformulierungen (vgl. z.B. Ash & Edgell 1975) oder Untersuchungen zur Unangenehmheit bestimmter Fragen in Subpopulationen (vgl. Koolwijk 1968) nützlich sein.

- j. Revision des Fragebogenentwurfs (h), evtl. auch der Entscheidung über die Befragungstechnik (g), die Fragestellung (b) bzw. die Untersuchungsvariablen (c) und Kovarianten (d) und erneute Erprobung des veränderten Fragebogenentwurfs. Veränderungen und erneute Erprobungen sind solange zu wiederholen, bis eine befriedigende ‚Endfassung‘ erstellt ist.
- k. Wahl eines angemessenen Verfahrens der Stichprobenziehung und Durchführung der Stichprobenziehung.
- l. Soweit persönliche Befragungen durchgeführt werden sollen, Auswahl und Schulung der Interviewer.
- m. Durchführung der Befragung, im Falle unpersönlicher (z.B. postalischer) Befragung mit mehrstufigem ‚Nachfassen‘.
- o. Formale Auswertung einschließlich statistischer Hypothesenprüfung.
Bei der Auswertung eines Fragebogens empfiehlt es sich, Plausibilitätskontrollen der Antworten durchzuführen, um Fälle zu entdecken, in denen z.B. die Eintragung der Antwort durch Interviewer oder Befragten an der falschen Stelle oder ohne Berücksichtigung des Inhalts nach einem bestimmten System erfolgt ist. Das gilt besonders beim Einsatz von Fragebogen zu diagnostischen Zwecken, da andernfalls gravierende Folgen für den Probanden eintreten können. Eintragungen an falscher Stelle lassen sich in der Regel natürlich nur entdecken, wenn richtige Antworten existieren.
- p. Interpretation der Ergebnisse unter Berücksichtigung der allgemeinen methodenspezifischen Beschränkungen bzw. Validitätsvorbehalte und gegebenenfalls auch der tatsächlich aufgetretenen methodischen Unzulänglichkeiten (z.B. Rücklauf bei postalischer Befragung).
- q. Einordnung der Befunde in den Wissensbestand, vor allem Darlegung von Abweichungen und Übereinstimmungen mit
 - fremden Befunden mit vergleichbarer Methode,
 - Befunden aufgrund andersartiger Methoden.
- r. Gegebenenfalls Erarbeitung von Hinweisen auf Probleme und Sachverhalte, die bei weiteren Untersuchungen innerhalb des Themenbereichs beachtet werden sollten.

Viele der o.a. Gesichtspunkte sind nicht für die Befragung als Untersuchungsmethode spezifisch oder berühren nicht die Konstruktion des Fragebogens im engeren Sinne. Ihre Auflistung macht deutlich, daß die Fragebogenkonstruktion nur ein (wenn auch wichtiger) Schritt im Zusammenhang mit einer empirischen Untersuchung oder diagnostischen Urteilsbildung ist. In den folgen-

den Abschnitten werden vor allem die bei der Erstellung des Fragebogenentwurfs erforderlichen Überlegungen zur Fragenformulierung, zur Festlegung der Fragenreihenfolge und zur äußeren Gestaltung des Fragebogens (Layout) eingehender behandelt. Ein Handbuchartikel muß sich dabei wegen des begrenzten verfügbaren Raumes mehr oder weniger auf eine Aufzählung von Problemen und Lösungsansätzen beschränken. Mindestens in einzelnen Aspekten weitergehende, teilweise auch stärker praxisbezogene Darstellungen sind unter vielen anderen Jonsson (1957), Noelle (1963), Richardson et al. (1965), Stroschein (1965), Oppenheim (1966), Phillips (1966, 1970), Richter (1969), Atteslander (1971), Friedrich (1971), Mayntz et al. (1971), Münch (1971), Friedrich (1973), Muccielli (1973), Scheuch (1973), Koolwijk & Wieken-Mayser (1974), Kreutz & Titscher (1974), Friedrich & Hennig (1975), Holm (1975b), Kirschhofer-Bozenhardt & Kaplitza (1975), Burisch (1976), Karmasin & Karmasin (1977). Über Befragungsmethoden allgemein informieren außerdem Jetzschmann & Kallabis (1966), Cannell & Kahn (1968), Anger (1969), König (1972), Maccoby & Maccoby (1972), Sheatsley (1972), Behrens (1974), Schreiber (1974), Steward & Cash (1974), Wilk (1974), Holm (1975a).

2. Fragentypen

2.1 Zielsetzungen von Fragen

Nicht jede in einem Fragebogen verwendete Frage hat die Aufgabe, inhaltlich oder im Sinne von ‚Zeichen‘ (signs) interessierende Informationen zu erheben. Bestimmte Fragen haben Merkmale im Auge, die nicht als solche interessieren und nur im Hinblick auf Interpretationen oder Erklärungen der eigentlich thematischen Sachverhalte von Bedeutung sind (dazu gehören in der Regel die Angaben zur Person des Interviewten). Außerdem gibt es Fragen, die nur innerhalb des Fragebogens bzw. der Befragung bestimmte Aufgaben zu erfüllen haben. In leichter Abwandlung der Terminologie vor Stroschein (1965) sei die erste Gruppe als die der ‚*Ergebnisfragen*‘, die zweite als die der ‚*Korrelationsfragen*‘ und die dritte als die der ‚*instrumentellen Fragen*‘ angesprochen.

Diese letztgenannte Gruppe, für die auch die Bezeichnung ‚*Funktionsfragen*‘ gebräuchlich ist (z.B. Anger 1969), läßt sich weiter unterteilen in:

- a. *Kontrollfragen*, und zwar einmal *Erhebungskontrollfragen* (z.B. Fragen nach Ort und Zeitpunkt des Interviews) zur Gewährleistung der Nachprüfbarkeit und *Auskunftskontrollfragen* (z.B. Wiederholungsfragen), mit dem Ziel der Ermittlung der Konsistenz des Antwortverhaltens,
- b. *Ablauf-Ordnungsfragen*, insbesondere *Filter-Fragen* mit der Aufgabe, Befragte mit bestimmten Merkmalen von bestimmten Fragen auszuschließen,

- und *Gablungs-* oder *Verzweigungs-Fragen* mit der Aufgabe, antwortabhängig zu verschiedenen Folgefragen zu verzweigen,
- c. *befragungstaktische Fragen* wie *Einleitungsfragen* mit der Aufgabe, den Kontakt zur Befragungsperson herzustellen, *Unterweisungsfragen* zur Information der Versuchsperson über evtl. nicht hinreichend bekannte Sachverhalte, *Ablenkungs-* und *Pufferfragen* zur Verdeckung der Zusammenhänge zwischen Fragen bzw. zur Vermeidung von Einflüssen vorangegangener auf nachfolgende Fragen (Halo-Effekte) und *Füllfragen* mit dem Ziel, dem Fragebogen in der Wahrnehmung des Befragten eine von der tatsächlichen abweichende inhaltliche Ausrichtung zu verleihen.

Erdos (1970) hält in postalischen Befragungen sogenannte *'return getters'*, also Fragen mit der alleinigen Aufgabe der Rücklaufsteigerung für angebracht. Für schwach- oder teilstandardisierte mündliche Befragungen werden z.B. von Stollberger (1966) und Atteslander (1971) außerdem *'Sondierungsfragen'* (das sind Nachfragen bei unzureichenden ersten Antworten) und *'Rangierfragen'* (Fragen, die im Falle von Abschweifungen den Befragten wieder auf das eigentliche Thema lenken sollen) als Typen taktischer Fragen angeführt. Weitere, zum Teil kurios anmutende Arten von Funktionsfragen beschreibt Noelle (1963, 74, sowie 1974, z.B. sogenannte Spielfragen nach der Beurteilung von Frisuren, Kleidern etc. mit dem Ziel, das Interesse der Vp am Interview zu erhalten).

Auch in diagnostischen Fragebogen wird von instrumentellen Fragen mehr oder weniger Gebrauch gemacht (vgl. Mittenecker 1971). So enthält etwa die englische Variante des MPI von Eysenck 12 Pufferfragen, die nicht in die Auswertung eingehen. Von den 566 Items des MMPI sind immerhin 166 instrumentelle, genauer *'Auskunfts-Kontroll-Items'*, die unterschiedliche Kontrollstrategien verfolgen: Itemwiederholungen zur Bestimmung der Konsistenz, Lügenitems (die wahrheitsgemäß nur in ganz bestimmter Richtung beantwortbar sind) und Sorgfaltsitems (die von fast allen Probanden in einer Richtung beantwortet werden, vgl. Hathaway & Mc Kinley 1963).

Kreutz & Titscher (1974) fordern, daß Itemwiederholungen nur in so großen Abständen erfolgen, daß die Antworten stochastisch voneinander unabhängig sind. Sie warnen - allerdings ohne dies empirisch zu begründen - vor den Folgen für die Motivation und das Antwortverhalten der Vpn, wenn diese die Wiederholung (und damit die Kontrollabsicht) bemerken.

Dem wäre entgegenzuhalten, daß die Entdeckung eingebauter Kontrollmechanismen durch die Vpn auch positive Wirkungen (größere Sorgfalt, größere Ehrlichkeit) haben kann. Nach den Ergebnissen von Hoeth & Köbler (1967) scheint es u.U. sogar vorteilhaft zu sein, wenn Vpn auf solche Mechanismen eigens hingewiesen werden (vgl. auch Ehlers 1973).

Ablauf-Ordnungsfragen (Filter- und Verzweigungsfragen) werden dagegen in diagnostischen Fragebogen nicht eingesetzt, da diese Fragebogen in der Regel schriftlich beantwortet werden und dadurch die Verwendung dieser Instrumente mit Schwierigkeiten verbunden wäre (Richter 1969). Überlegungen zu individualisiertem (antwortabhängigem) Testen im Bereich der Persönlichkeitsdiagnostik - allerdings unter Aufgabe des traditionellen Fragebogenkonzeptes - sind damit natürlich nicht ausgeschlossen.

2.2 Frageninhalte

Als mögliche Inhalte von Ergebnisfragen (vgl. 2.1) kommen vor allem in Betracht (Stollberger 1966, Holm 1975b):

- Fakten (z.B. Lebensalter, Besitz eines Farbfernsehgerätes).
- Wissen („Wie heißt der Bundesfinanzminister?“). Während bei Faktenfragen das Interesse des Untersuchers sich auf das Faktum richtet, d.h. er etwas über das Faktum (Verbreitung von Farbfernsehgeräten) erfahren will, interessiert bei Wissensfragen letztlich nicht das Faktum (der Name des Politikers), sondern die Informiertheit des Befragten.
- Beurteilungen, Bewertungen, Meinungen bzw. Einstellungen („Was halten Sie von Kernkraftwerken?“). Soweit der Untersucher an Informationen über Sachverhalte (Kernkraftwerk) interessiert ist, spricht man von Beurteilungs- oder Bewertungsfragen, steht dagegen der Befragte im Mittelpunkt des Interesses, von Meinungs- bzw. Einstellungsfragen (Holm 1975 b).
- Verhalten bzw. Handlungen („Treiben Sie regelmäßig Sport?“). Dabei handelt es sich nur dann um Faktenfragen, wenn nach gegenwärtigem oder früherem Verhalten, nicht aber wenn nach zukünftigem oder hypothetischem Verhalten gefragt wird.
- Motive („Warum sind Sie dieser Meinung?“).

Nach Cannell et al. (1977) ist davon auszugehen, daß unabhängig von der Fragenform Einstellungs- bzw. Motiv- im Vergleich zu Fakten- oder Wissensfragen ‚schwieriger‘ sind: Sie führen häufiger zu ausweichenden Antworten („weiß nicht“ o.ä.), häufiger zu Rückfragen das Fragenverständnis betreffend und häufiger zu qualifizierten (eingeschränkten) Antworten.

Das bedeutet allerdings nicht, daß Fakten- oder Wissensfragen auch zu ‚richtigeren‘ Antworten führen müßten bzw. daß es einfacher sei, Wissen und Fakten durch geeignete Fragen zu erfassen (Mauldin & Marks 1950).

Eine Grundforderung an Fragen ist die nach Eindeutigkeit in einem gegebenen Zusammenhang. Daraus ergibt sich, daß eine Frage sich stets nur auf einen bestimmten Inhalt beziehen darf. In diesem Sinne wäre z.B. das Item (MMPI Nr. 307) „Bei einigen Spielen lehne ich es ab, mich zu beteiligen, weil ich sie

nicht gut kann. richtig/falsch?' nicht eindeutig, da hier gleichzeitig nach einem Faktum und nach einem Motiv gefragt wird. (Trotz dieses inhaltlichen Mangels ist nicht ausgeschlossen, daß im Rahmen eines empirischen Validitätskonzeptes ein solches Item sich als brauchbar erweist.)

2.3 Direktheit einer Frage

Neben der direkten Frage (‚Wie alt sind Sie?‘) zu einem bestimmten Sachverhalt gibt es meist mehrere Möglichkeiten zur Formulierung indirekter Fragen, die nach verschiedenen Gesichtspunkten eingeteilt werden können. Holm (1974a) bezeichnet als indirekte Fragen solche, die sich direkt auf einen verwandten Sachverhalt beziehen, besonders

- ‚Fragen durch die Hintertür‘ (‚Welcher Jahrgang sind Sie?‘),
- ‚Fragen über Ersatzdimensionen‘ (‚Wieviele Zähne fehlen Ihnen?‘).

Stroschein (1965), der von ‚unmittelbaren‘ und ‚mittelbaren‘ Fragen spricht, führt als Beispiele an:

- Assoziationsfragen, d.h. Fragen, die darauf abzielen, die mit einem bestimmten Gegenstand verbundenen Vorstellungen zu ermitteln,
- Projektionsfragen, d.h. Fragen, die den Probanden veranlassen sollen, in eine Situation oder Person eigene Gefühle oder Stimmungen hineinzuverlagern.

Weiterhin werden ‚Dialogfragen‘ (Stroschein 1965, ‚A sagt . . . , B sagt. . . .‘) häufig so gestellt, daß der Befragte seine Meinung indirekt in Form eines Schiedsspruches (‚Wer hat recht?‘) zwischen A und B zum Ausdruck bringen soll. Ähnliches gilt für sogenannte ‚hypothetische Situationen‘ (‚Stellen Sie sich bitte vor, Herr X‘), die prinzipiell in eine direkte (‚Wie würden Sie sich verhalten?‘) oder eine indirekte Frage (‚Wie glauben Sie, daß Herr X sich verhält?‘) münden können (vgl. Friedrichs 1973). In gewisser Hinsicht als indirekt muß man wohl auch ‚Fragen mit Zitaten‘ ansehen, bei denen eine bestimmte Meinung vom Probanden nicht unmittelbar, sondern über den Umweg der Stellungnahme zu einem Zitat erfragt wird, das der Untersucher einer in der Regel bekannten Persönlichkeit in den Mund legt. Wie auch empirisch vielfach demonstriert wurde, hängt die Antwort außer vom Inhalt des Zitates natürlich stark von der Einstellung des Probanden gegenüber der zitierten Persönlichkeit ab (vgl. z.B. Roslow et al. 1940, Stroschein 1965). Weitere Beispiele indirekter Fragen finden sich u.a. bei Karmasin & Karmasin (1977) und bei Maccoby & Maccoby (1972).

Indirekte Fragenformulierungen werden meist für sogenannte schwierige, heikle, unangenehme, peinliche Sachverhalte verwendet. Dazu gehören für viele Befragte Fragen nach dem Einkommen, der Kindererziehung, der Allge-

meinbildung, der Sexualität, den Familienverhältnissen (Friedrichs 1973) und vor allem nach der körperlichen Sauberkeit (Scheuch 1973). Aber auch für die Leserschafts-Forschung werden indirekte Techniken vorgeschlagen (vgl. z.B. Schyberger 1968), um Verfälschungen durch die Auflagenstärke zu vermeiden. Barton (1958) hat ironisierend am Beispiel der (direkten) Frage ‚Haben Sie Ihre Frau umgebracht?‘ die verschiedenen mehr oder weniger indirekten Ansätze zusammengestellt:

- die Möglichkeitsfrage (‚Könnte es sein, daß . . .‘),
- die Kartenfrage (Identifizierung der Kennung einer Karte mit der zutreffenden Antwort durch die VP),
- der Jedermann-Ansatz (‚Viele haben in letzter Zeit . . . und Sie?‘),
- der Andere-Ansatz (‚Kennen Sie Leute, die . . . und Sie?‘),
- die Urnentechnik (Antwort auf direkte Frage kommt in verschlossenem Umschlag in eine Urne),
- die projektive Technik (‚Welche Gedanken kommen Ihnen bei diesen Bildern . . .?‘),
- die ‚Kinsey-Technik‘ (dem evtl. peinlichen Verhalten wird durch die Formulierung die Eigenschaft des Selbstverständlichen verliehen).

Inwieweit indirekte Fragen, vor allem Projektions- und Assoziationsfragen, die in sie gesetzten Erwartungen erfüllen, scheint weitgehend ungeklärt zu sein (Stroschein 1965, Friedrichs 1973).

Karmasin & Karmasin (1977) bemerken kritisch, daß solche Fragen für die Vpn grundsätzlich mehrdeutig seien. Die Vp kann bei einer indirekten Frage (vgl. o. ‚Wer hat recht?‘)

- ihre eigene Meinung äußern,
- sich überlegen, wie das ‚mehrheitlich‘ wohl gesehen wird, oder
- unter Zugrundelegung irgendwelcher Normen (wie es sein sollte) entscheiden.

Die Interpretation der Antwort auf eine indirekte Frage ist demnach nur auf dem Hintergrund einer Theorie möglich, die den Zusammenhang zwischen Sachverhalt und Frage herstellt (z.B. eine Theorie der Projektion, vgl. Anger 1969, aber auch ein empirisches Validitätskonzept). Dabei ist die Indirektheit einer Frage als Kontinuum anzusehen. Der Grad der Indirektheit bestimmt sich nach der Komplexität der Mechanismen, die die Theorie zur Vermittlung zwischen Sachverhalt und Antwort annimmt (Cannell & Kahn 1968). Indirekte Fragen sind demnach höchstens so brauchbar, wie es die ihnen zugrundeliegende Theorie ist.

Für Items in diagnostischen Fragebogen hat Ellis (1947) in einer umfangreichen Untersuchung die direkte (hier personalisierte) Formulierung (‚Ich . . .‘) mit der indirekten (unpersönlichen, ‚Leute, die . . . sind . . .‘) verglichen. Dabei

zeigte sich für die indirekte (unpersönliche) Form eine stärkere Abhängigkeit von anderen Fragebogenmerkmalen (positive oder negative Formulierung) als für direkte (personalisierte) Items. Vor allem aber war auch bei heiklen (sozial mutmaßlich unerwünschten) Sachverhalten keine Überlegenheit der indirekten Formulierung nachweisbar.

2.4 Formale Fragenkonstruktion

Prinzipiell kann ein Fragebogenitem als Frage („Besitzen Sie ein Farbfernsehgerät? ja/nein“) oder als Statement mit Aufforderung zur Stellungnahme („Ich besitze ein Farbfernsehgerät. stimmt/stimmt nicht“) formuliert sein. In der Literatur über Fragebogenkonstruktion scheint „... der formale Unterschied zwischen Behauptungen (Statements) und Fragen noch gar nicht recht bewußt geworden ...“ zu sein (Kreutz & Titscher 1974, 52). Über Auswirkungen dieses Unterschiedes läßt sich deshalb derzeit nur spekulieren. Kreutz & Titscher (1974) vermuten z.B., daß die in vielen diagnostischen Fragebogen beobachtete Zustimmungstendenz (acquiescence) mit der üblichen Formulierung der Items als Statements zusammenhängt und sich weniger ausgeprägt zeigen würde, hätten diese als Fragen in stärkerem Maße den Charakter des „Unentschiedenen“.

2.4.1 Offene und geschlossene Fragen

Ausgehend von den Antwortmöglichkeiten auf eine Frage unterscheidet man *offene* von geschlossenen Fragen (z.B. Stollberger 1966, Friedrichs 1973, Cannell et al. 1977). Eysenck (1953) spricht vom „kreativen“ und vom „selektiven“ Antworttyp, und Stroschein (1965) hebt die „inkategorialen Fragen“, d.h. die Fragen, bei denen die Auswertungsgesichtspunkte für die Vpn nicht erkennbar sind, von den „kategorialen Fragen“ ab, die in irgendeiner Form (meist durch Antwortvorgaben) Informationen über die Auswertungsgesichtspunkte enthalten.

Geschlossene Fragen lassen sich „kategorie-neutral“ (Stroschein 1965), d.h. ohne durch die Vorgabe der Kategorien das Antwortverhalten gravierend zu beeinflussen, zu einem bestimmten Themenbereich nur formulieren, wenn alle möglichen Antworten bereits bekannt sind. Aus diesem Grund wird man bei geringem Vorwissen zwangsläufig eher zu offenen Fragen greifen (Friedrichs 1973). Daneben ist aber auch zu bedenken, daß offene Fragen „freie Reproduktion“, geschlossene Fragen nur „Wiedererkennen“ fordern. Schon dadurch sind offene Fragen schwieriger. Cannell et al. (1977) berichten von ca. 30% unbrauchbaren Antworten (z.B. Auslassungen und nicht fragenbezogene Ausführungen) bei offenen im Vergleich zu nur 6% bei geschlossenen Fragen

zum gleichen Themenkomplex, merken allerdings selbstkritisch an, daß man der geschlossenen Antwort ihre Unbrauchbarkeit oft auch nur nicht ansieht. In die Richtung größerer Schwierigkeit der offenen Frage deuten aber (abgesehen von Plausibilitätsüberlegungen hinsichtlich erforderlicher Ausdrucksfähigkeit und evtl. Schreibgewandtheit) auch Ergebnisse über Rückläufe bei postalischen Befragungen: Falthzik & Carroll (1971) erzielten bei einem aus nur einer Frage bestehenden Fragebogen einen Rücklauf von 78%, wenn die Frage geschlossen formuliert war, und von nur 27%, wenn es sich um eine offene Frage handelte. Einen Unterschied von immerhin noch 60% zu 50% berichtet Erdos (1970). Andererseits fand Richter (1969) für umfangreichere Fragebogen zwar ebenfalls eine Senkung des Rücklaufs durch eine große Zahl offener Fragen, stellt aber günstige Auswirkungen auf den Rücklauf fest, wenn zu jedem Themenbereich neben geschlossenen Fragen auch eine offene Frage vorgesehen ist, was er auf eine ‚Ventilfunktion‘ offener Fragen und auf verminderte ‚Ermüdung‘ durch vereinzelte Zwischenschaltung solcher Fragen zurückführt.

Zu bedenken ist auch, daß die ‚Abdeckung‘ eines Themenbereiches in der Regel erheblich mehr geschlossene als offene Fragen erfordert (Cannell et al. 1977).

Letztlich wird aber entscheidend für die Wahl offener oder geschlossener Fragen sein, ob die Reproduktions- oder die Wiedererkennungsleistung dem untersuchten Inhalt angemessener ist. So demonstrierten z.B. Roslow et al. schon 1940, daß bei der Ermittlung von Kaufgewohnheiten und Verbreitungsgraden in offenen Fragen (freie Reproduktion) häufiger als in geschlossenen Fragen (Wiedererkennen) und auch häufiger als objektiv zutreffend die Produkte mit hohen Marktanteilen genannt wurden (ähnliche Ergebnisse berichtet Stroschein 1965). Wie genau hier die geschlossene Frage den wahren Sachverhalt trifft, hängt entscheidend von der Vollständigkeit der Vorgaben ab (Roslow et al. 1940). Ungeeignet sind offene Fragen auch zur Erfassung ‚alltäglicher‘ Sachverhalte (Payne 1951), die in der Regel unscheinbar, d.h. nicht als Figur abgehoben sind (auf eine offene Frage nach dem Tagesablauf hin werden viele Vpn z.B. das ‚Zähneputzen‘ nicht erwähnen).

Dagegen ist die offene Frage wegen geringen Vorwissens, hoher Differenziertheit bzw. Komplexität des Sachverhaltes (Friedrichs 1973), hoher problemspezifischer Validität der Reproduktionsleistung oder Unangenehmheit bzw. SD-Empfindlichkeit des Inhalts (Sudman & Bradburn 1974, Bradburn & Sudman 1979) in anderen Fällen durchaus angezeigt, allerdings bereitet ihre Auswertung erhebliche Schwierigkeiten. Die freien Antworten werden üblicherweise nach Art einer systematischen Inhaltsanalyse (vgl. z.B. Friedrichs 1973) mit Hilfe eines eigens erstellten Kategoriensystems klassifiziert (grundsätzliche Überlegungen speziell für den Fall offener Fragen finden sich z.B. bei Lazarsfeld & Barton 1955, Vorschläge zur ‚Automatisierung‘ unter Einsatz von EDV

berichten Friesbie & Sudman 1968). Dabei zeigt sich allerdings häufig, daß relativ viele Antworten (mindestens 10%) bei Zugrundelegung eines noch überschaubaren Systems nicht klassifizierbar sind und daß intraindividuelle Stabilität und interindividuelle Objektivität der Klassifizierung zu wünschen übrig lassen (Stroschein 1965). Außerdem führen vorgegebene Antwortkategorien und Klassifikationen freier Antworten unter Zugrundelegung dieser ‚Antwortvorgaben‘ häufig zu recht unterschiedlichen Ergebnissen (Stroschein 1965), was sich allerdings ebensogut als Argument gegen offene wie gegen geschlossene Fragen verwenden läßt.

Die Unterscheidung zwischen offenen und geschlossenen bzw. inkategorialen und kategorialen Fragen ist entgegen dem ersten Anschein nicht streng durchzuführen (Maccoby & Maccoby 1972). Einmal ist es möglich, in der Fragenformulierung nur bestimmte Antwortkategorien aufzuführen, aber klarzumachen, daß es weitere gibt (‚Haben Sie gestern abend ferngesehen oder . . . ?‘). Noelle (1963) spricht dann von ‚halboffenen‘ Fragen. Zum anderen kann eine Frage auch nur ‚scheinbar inkategorial‘ (Stroschein 1965) sein, weil die möglichen Kategorien jedem Befragten evident sind (das gilt z.B. für die ‚offene‘ Frage: ‚Welche Haarfarbe haben Sie?‘).

Im Falle mündlicher Befragung muß schließlich berücksichtigt werden, daß eine Frage aus der Sicht des Probanden offen sein, die Antwort vom Interviewer aber direkt klassifiziert werden kann. Gegen diese sogenannte ‚Feldverschlüsselung‘ (vgl. auch Noelle 1963) werden allerdings gravierende Einwände erhoben (Anger 1969), da sie den Interviewer häufig überfordere und vor allem nicht nachprüfbar sei. Andererseits werden auch bei der schriftlichen Erfassung freier Antworten durch Interviewer erhebliche ‚Verluste‘ beklagt, so daß Quantifizierungen mit Vorsicht behandelt und Hypothesenprüfungen auf der Grundlage offener Fragen nicht vorgenommen werden sollten (Anger 1969).

Da offene Fragen im Rahmen eines mündlichen Interviews eher der alltäglichen Konversation entsprechen und daher natürlicher wirken (Maccoby & Maccoby 1972, Karmasin & Karmasin 1977), werden sie auch aus vorwiegend befragungstaktischen Gründen eingesetzt.

2.4.2 Arten geschlossener Fragen

Streng kategorial sind Fragen, wenn sie explizit alle Antwortmöglichkeiten enthalten. Bei der Formulierung ist im allgemeinen sicherzustellen, daß die Fragen kategorie-neutral sind, d.h. daß keine der Antwortkategorien durch die Formulierung begünstigt wird.

Neutralität im Hinblick auf die *Reihenfolge* der Vorgaben ist allerdings nur erreichbar, wenn mit verschiedenen Fassungen des Fragebogens gearbeitet und

dabei die Reihenfolge der Vorgaben variiert wird (Split-ballot-verfahren, gega-belte Befragung).

Zur Vermeidung von Einflüssen z.B. der sozialen Erwünschtheit kann es in Ausnahmefällen angezeigt sein, bewußt und geplant von der Ausgewogenheit der Kategorien abzuweichen. So berichtet z.B. Stroschein (1965), daß die Frage ‚Werden Sie bestimmt mitwählen oder werden Sie vielleicht nicht zur Wahl gehen?‘ zu einer sehr guten Prognose der Wahlbeteiligung führte, vermutlich gerade weil sie nicht kategorie-neutral ist (vgl. auch 3.1.4).

Nach der Zahl der Antwortvorgaben lassen sich kategoriale Fragen weiter unterteilen in *Alternativfragen* und in *Selektivfragen* (Listenfragen, Auswahl-fragen, auch Katalogfragen, vgl. Anger 1969, oder Multiple-Choice-Fragen, vgl. Payne 1951, obschon der letztgenannte Begriff vorwiegend bei Verwen-dung entsprechender Fragen im Rahmen diagnostischer Fragebogen gebraucht zu werden scheint). Ein Spezialfall sind sogenannte Eigenschaftswörterlisten (adjective check lists). Dabei ist zu berücksichtigen, ob

- die Zahl zulässiger Nennungen unbestimmt bleiben,
- die Zahl zulässiger Nennungen nach unten und/oder nach oben begrenzt werden oder ob
- zu jeder der aufgeführten Kategorien eine Einzelantwort gefordert werden soll.

Problem der Alternativ-, vor allem der Listenfragen ist die Bevorzugung von Vorgaben in Abhängigkeit von ihren Positionen (darauf wird in 3.1.3 näher eingegangen).

Selektivfragen mit vielen Vorgaben sind schwierige Fragen (nach Richter 1969 senken sie den Rücklauf in postalischen Befragungen) und machen im mündli-chen Interview Hilfsmittel (Vorlagen, Kartensätze) erforderlich. Werden sol-che nicht verwendet, erweisen sich umfangreiche Selektivfragen als besonders anfällig gegenüber Interviewereinflüssen (Cahalan et al. 1947).

Sonderfälle selektiver Fragen sind sogenannte *Skalafragen* (ordinale oder im engeren Sinne quantitative Fragen, vgl. Holm 1975 b), d.h. Fragen, deren Antwortkategorien geordnet bzw. graduell abgestuft sind. Verwendet man solche Fragen im Zusammenhang mit Beurteilungen, Bewertungen oder Ein-schätzungen, so spricht man auch von *Ratingskalen*. über diesen Ansatz orientieren zusammenfassend Guilford (1954) und Clauss (1968), neuere Er-gebnisse zur Frage der optimalen Zahl von Skalenstufen referiert Mc Kelvie (1978), Varianten des innerhalb demoskopischer Befragungen besonders be-liebten graphischen Rating diskutiert Narayana (1977). ‚Geeichte‘ numerisch-verbale Skalen für Häufigkeiten, Intensitäten, Wahrscheinlichkeiten und Beur-teilungen des Zutreffens finden sich bei Rohrmann (1978).

Ratingskalen werden z.B. auch zur Erfassung des Zustimmungsgrades in diagnostischen Fragebogen verwendet, wobei die zweistufige Variante (ja/nein; richtig/falsch) wiederum als Spezialfall angesehen werden kann. Die ebenfalls verbreitete dreistufige Form (ja/?/nein) bietet zwar die Möglichkeit, über die Häufigkeit von ‚?‘ - Antworten die Aktualität der Fragen für den Probanden zu ermitteln (Heller & Krüger 1976), wirft andererseits aber wie alle mehrstufigen Kategorienskalen Probleme im Hinblick auf die Dimensionalität der Antworten auf (vgl. 1.1.3).

2.4.3 Sonderformen

Soweit nicht ohnehin schriftliche Befragung erfolgt, empfiehlt sich bei umfangreichen Selektivfragen die Verwendung von Vorlagen, entweder als

- Einzelvorlage mit allen Kategorien oder als
- Vorlagensatz (Kartensatz) mit je einer Vorlage für jede Kategorie.

Die Verwendung von Kartensätzen hat die Vorteile der leichten Variierbarkeit der Reihenfolge und der offenbar größeren Sorgfalt der Vpn bei der Entscheidung (Stroschein 1965). Vor allem werden Karten an späterer Stelle stärker beachtet als Vorgaben auf den unteren Plätzen einer Liste.

Zur Kennzeichnung der Vorgaben kann sich die Verwendung von *Symbolen* empfehlen (a, b . . . ; weiß, schwarz . . . ; 1, 2 . . . etc). Dadurch lassen sich Übertragungsfehler bei der Kornunikation Interviewer/Befragter vermindern. Vor allem aber wird der Befragte der Notwendigkeit enthoben, die Antwort explizit auszusprechen, was ihm bei unangenehmen Fragen peinlich sein könnte. Entgegen der landläufigen Erwartung haben sich nach Stroschein (1965) keine Antwortbevorzugen durch bestimmte zur Kennzeichnung verwendete Symbole nachweisen lassen. Interviewereinflüsse auf die Antworten scheinen bei Verwendung von Symbolen geringer zu werden.

Als weitere Sonderform wäre das ‚semantische Differential‘ (Osgood et al. 1957) oder ‚Polaritätenprofil‘, eine Zusammenstellung von meist 18 bipolaren ‚Dimensionen‘ (adjektivischen Gegensatzpaaren, die jeweils durch eine siebenstufige Skala miteinander verbunden sind), anzuführen. Die Einschätzung von Begriffen in diesen 18 Polaritäten läßt sich - relativ unabhängig von den konkret verwendeten Eigenschaftspaaren - als Lokalisation dieser Begriffe in einem semantischen Raum mit den Dimensionen der Bewertung (gut/schlecht), der Aktivität (aktiv/passiv) und der Intensität (stark/schwach) interpretieren (vgl. auch Herrmann & Stäcker 1969).

Daneben werden vor allem im Bereich der kommerziellen Markt- und Meinungsforschung zahlreiche weitere, meist als ‚psychologisch‘ bezeichnete Techniken (von Farbwahltests bis zum Baumtest, vgl. Noelle 1963) unkritisch

und in einer Weise eingesetzt, die in erstaunlichem Kontrast zu den in anderem Zusammenhang (z.B. bei der Standardisierung der Fragenformulierungen und bei der Auswertung) erhobenen Forderungen nach Objektivität und Nachprüfbarkeit steht. Die nach Anger (1969, 583) „dringend benötigte Information über wichtige individuelle Merkmale und Eigenschaften“ läßt sich, soweit es sich um psychische Eigenschaften handelt, nicht mittels irgendwelcher ‚Kurzverfahren‘ durch wenig geschultes Personal (Interviewer) nebenbei beschaffen. Den diesbezüglichen Ausführungen und Empfehlungen von Anger (1969) muß mit erheblicher Skepsis begegnet werden.

3. Fragenformulierung

Die Formulierung von Fragen wird von den meisten Autoren als ‚Kunst‘ betrachtet (u.v.a. Noelle 1963, Mayntz et al. 1971, Scheuch 1973), deren Grundlagen nicht am Schreibtisch, sondern nur in langer Erfahrung erworben werden könnten. Demgemäß sind auch ‚Regeln‘ für die Formulierung von Fragen, wie sie in der Literatur vielfach angeführt werden (z.B. Payne 1951, Edwards 1957, Noelle 1963, Maccoby & Maccoby 1972), soweit sie konkret sind, bestenfalls Ausfluß solcher Erfahrungen (oder wie im Falle von Holm 1975 b problematischer theoretischer Ansätze) und unbewiesen oder so abstrakt, daß sie zwar mit hoher Wahrscheinlichkeit nicht falsch, aber dafür auch wenig hilfreich sind (Kreutz & Titscher 1974). Die folgenden Ausführungen beinhalten weniger eine (erneute) Wiedergabe solcher ‚Regeln‘ als eine Beleuchtung grundsätzlicher Probleme und eine Zusammenstellung empirischer Befunde, die sich naturgemäß nur beschränkt verallgemeinern lassen.

Eine Frage stellt einerseits einen verbalen Stimulus für den Befragten, andererseits ein sprachliches Abbild eines Sachverhaltes dar (Kreutz & Titscher 1974). Bevor ein Sachverhalt sprachlich abgebildet werden kann, muß seine inhaltliche Struktur festliegen, d.h. der sprachlichen Formulierung einer Frage geht logisch die Erarbeitung einer inhaltlichen Konzeption voraus. In der Praxis allerdings lassen sich beide Schritte nicht wie hier mehr oder weniger streng trennen, mitunter ist sogar schwer zu entscheiden, ob es sich bei einem konkreten Problem um ein vorwiegend inhaltliches oder vorwiegend sprachliches handelt (vgl. z.B. die affektiv nicht neutralen Begriffe). Im Zusammenhang mit der Entwicklung der inhaltlichen Fragenkonzeption ist auch die Entscheidung über den zur Verwendung kommenden Fragentyp zu treffen. Die dabei zu beachtenden Gesichtspunkte sind bereits im vorstehenden Kapitel behandelt worden.

Wesentliche empirische Befunde zur Auswirkung der Fragenformulierung auf Antworten stammen aus der Zeit des Zweiten Weltkriegs, später traten Interviewereinflüsse in den Vordergrund des Interesses (Hartmann 1972).

3.1 Die inhaltliche Konzeption einer Frage

Ist bei der Planung einer Befragung (vgl. 1.3) die abhängige Variable (Programmfrage im Sinne von Noelle 1963) definiert worden (z.B. ‚Beurteilung der Wirtschaftspolitik der Bundesregierung durch die Bevölkerung‘), so muß sie in einem weiteren Schritt operationalisiert, d.h. in eine oder mehrere Erhebungs- bzw. Testfragen (Noelle 1963) übersetzt werden.

3.1.1 Vorüberlegungen

Zunächst ist in Abhängigkeit vom Forschungsziel bzw. von der Programmfrage festzulegen, ob eine *globale* oder *differenzierte* Vorgehensweise oder eine Kombination beider erfolgen soll. Payne (1951) zeigt auf, daß z.B. bei Beurteilungen das aus mehreren Urteilen über Einzelaspekte sich ergebende Bild oft nicht mit dem einer global erfolgten Beurteilung übereinstimmt. Sodann ist zu klären, ob es sich um eine

- normative oder deskriptive,
- kognitive oder evaluative,
- allgemeine oder spezifische,
- abstrakte oder konkrete

Frage handeln soll (Karmasin & Karmasin 1977). Dabei sind allerdings die Freiheitsgrade des Untersuchers eingeschränkt, z.B. ist zu berücksichtigen, daß statt einer vorgesehenen kognitiven Beurteilung bei fehlender Informationsbasis auf Seiten der Vp leicht eine Bewertung (Evaluation) zustande kommen kann (z.B. bei einer Frage über Auswirkungen der Hochzinspolitik). Andererseits ist es sicher auch problematisch, die Menge der zulässigen Fragenkonzeptionen von vornherein stark einzuschränken, etwa auf spezifische und konkrete Ansätze (z.B. im Sinne von Payne 1951, der allgemein einen Bezug der Frage auf das ‚Wer, Wann, Warum, Wo, Wie‘ fordert, vgl. auch Anger 1969, Friedrichs 1973). So betonen Karmasin & Karmasin (1977), daß es zahlreiche Sachverhalte gebe, bei denen das ‚Dogma‘ von der konkreten und spezifischen Frage zu unsinnigen Konsequenzen führen müßte. Bei der Untersuchung von Lesegewohnheiten z.B. interessiert durchaus nicht, wie lange eine bestimmte Zeitung an einem bestimmten Tag gelesen wurde (spezifischer und konkreter Ansatz), sondern wie lange (ausführlich o.ä.) ‚in der Regel‘ die jeweilige Tageszeitung gelesen wird (allgemeiner, abstrakter Ansatz).

Soweit von der Vp Gedächtnisinhalte abgerufen werden müssen, ist zu überlegen, wie mit einer in der Regel identischen Frage die möglicherweise sehr unterschiedlichen Erfahrungen verschiedener Vpn aktualisiert werden können (Cannell & Kahn 1968) und wie eine für die Reproduktion bzw. das Wiedererkennen optimale, an ‚cues‘ reiche Situation hergestellt werden kann (Cannell et al. 1977, vgl. auch 1.2.3).

Für Beurteilungen ist festzulegen, ob sie auf dem Hintergrund eines impliziten Bezugssystems der Vp erfolgen oder - soweit ein solches nur unvollkommen ausgebildet oder interindividuell stark unterschiedlich ist - durch vergleichende Urteile vorgenommen werden sollen (vgl. Payne 1951, die ‚absolute‘ Beurteilung des Nährwertes von Milch führt vermutlich zu wenig brauchbaren Ergebnissen; es bietet sich an, Vergleiche mit anderen Nahrungsmitteln vornehmen zu lassen).

Damit eine Frage eindeutig ist, darf sie nur einen relevanten Gesichtspunkt (eine ‚Dimension‘) enthalten. Bei der Entwicklung der inhaltlichen Fragenkonzeption muß man also einerseits sicherstellen, daß der gewünschte Aspekt erfaßt ist, gleichzeitig müssen andere Aspekte ausgeschlossen sein. Das berühmte Beispiel einer vieldimensionalen Frage von Lazarsfeld (1935): ‚Warum haben Sie dieses Buch gekauft?‘ (Dimensionen können u.a. sein: ‚Sie‘ vs. andere Menschen, ‚dieses‘ vs. andere Bücher, ‚Buch‘ vs. andere Gegenstände, ‚gekauft‘ vs. andere Formen des Erwerbs) ist eine von der Konzeption her unangemessene Frage, da zwar die interessierende Dimension enthalten ist, andere aber nicht ausgeschlossen wurden. Neben diesen ersten allgemeinen Überlegungen zur Fragenkonzeption sind mehrere spezielle Gesichtspunkte zu berücksichtigen. Sie werden in den nachfolgenden Abschnitten behandelt.

3.1.2 Definition des Gegenstandes und Explikation eines Bezugsrahmens

In aller Regel kann nicht davon ausgegangen werden, daß die Vpn über eine einheitliche und mit der des Untersuchers übereinstimmende Vorstellung vom Befragungsgegenstand verfügen. So berichtet Noelle-Neumann (1970), daß die Frage nach dem Besitz bzw. der Verwendung einer ‚Perücke‘ von 1% der Befragten bejaht wurde, wenn gleichzeitig auch nach einem ‚Haarteil‘ gefragt wurde (15% Ja-Antworten). Wurde dagegen ohne weitere Unterscheidung nur nach einer ‚Perücke‘ gefragt, antworteten 8% der Befragten mit ‚Ja‘. Dies läßt sich wohl nur so erklären, daß für einen Teil der Befragten auch Haarteile zu ‚Perücken‘ gehören, für einen anderen Teil dagegen nicht. Es ist also erforderlich, entweder eine möglichst exakte Definition für den Befragungsgegenstand vorzugeben oder aber - was im Sinne der Untersuchungsfragestellung ebenfalls interessant sein kann - die bei der Antwort zugrunde gelegte Definition von den Vpn zu erfragen (Cannell & Kahn 1968). Entsprechendes gilt für das Bezugssystem von dem ausgehend die Vpn ihre Antworten formulieren und das interindividuell sehr unterschiedlich sein kann (sicherlich haben Texaner und Bewohner Alaskas unterschiedliche Vorstellungen davon, was ein ‚warmer Sommer‘ ist, Cannell & Kahn 1968). Dabei muß das Bezugssystem - sofern es vorgegeben und nicht erfragt wird - für die Befragten relevant sein: Würde man eine Hausfrau nach dem jährlichen Eierverbrauch befragen, würde ihre Antwort reines Raten sein, der relevante Zeitraum hier z.B. ist die Woche

(Payne 1951). Darüber hinaus ist es u.U. erforderlich, eine Skala (Absolutwerte, Prozentangaben) für die erfragten Quantitäten zu spezifizieren und die erforderliche Genauigkeit festzulegen, um die üblicherweise beobachteten Antworthäufungen bei runden Zahlen zu vermeiden (Payne 1951). Besonderes Gewicht muß bei offenen Fragen auf die Definition von Befragungsgegenstand und Bezugssystem gelegt werden, da diese Fragen nicht durch explizite Antwortvorgaben eine weitere Einengung und Festlegung erfahren (Anger 1969).

3.1.3 Festlegung der Antwortkategorien

Soll eine Frage kategorie-neutral (Stroschein 1965, vgl. auch 2.4.1) sein, d.h. nicht schon durch die Vorgabe der Antwortkategorien bestimmte Antworten begünstigen (zu sogenannten ‚verzerrten‘ Fragen - Friedrichs 1973 - und ihrer legitimen Verwendung vgl. 3.1.4), so müssen die Vorgaben erschöpfend und - falls Mehrfachnennungen nicht erlaubt sind - disjunkt sein (Beispiele für Probleme bei nicht erschöpfenden Vorgaben finden sich bei Payne 1951, 87). Im Text der Frage müssen alle Vorgaben genannt werden oder es darf keine Vorgabe enthalten sein (Noelle-Neumann 1970). Zulässige Abweichungen von diesem Grundsatz beschreiben Kreutz & Titscher (1974).

Ausgewogen sind Vorgaben dann, wenn sie zu gleichen Teilen und mit gleicher Gewichtigkeit ‚positive‘ und ‚negative‘ Äußerungen dem jeweiligen Sachverhalt gegenüber beinhalten (Verstöße gegen diese Forderung und ihre Auswirkungen auf die Antworten beschreiben z.B. Payne 1951 und Rugg & Cantril 1972). Bei der Formulierung der Antwortkategorien ist auf möglichst vergleichbare soziale Wünschbarkeit zu achten (Phillips 1966) und zu bedenken, daß sich die ‚Attraktivität‘ erheblich durch den Aufweis von Konsequenzen der Antwort beeinflussen läßt (z.B. erfährt eine vorgeschlagene Rentenerhöhung erheblich weniger Zustimmung, wenn die entsprechende Antwortvorgabe auch die Konsequenz einer Erhöhung der Beiträge zur Rentenversicherung deutlich macht, vgl. Karmasin & Karmasin 1977). Dies gilt verstärkt, wenn diese Konsequenzen personalisiert werden (‚. . . wenn Sie dafür einen höheren Beitrag zur Rentenversicherung zahlen müßten . . .‘ vs., ‚. . . wenn dadurch die Beiträge zur Rentenversicherung steigen würden . . .‘), wie überhaupt die Personalisierung von Fragen deutliche Veränderungen im Antwortverhalten zur Folge haben kann (Rugg & Cantril 1972).

Daneben spielt aber auch die ‚Extremheit‘ der verwendeten Vorgaben eine erhebliche Rolle: Karmasin & Karmasin (1977) zeigen am Beispiel zweier Befragungen zur gesetzlichen Regelung des Schwangerschaftsabbruchs, wie durch Hinzufügung einer extremeren Antwortkategorie (völlige Freigabe) die Befürwortung der Fristenlösung erheblich zunimmt. Beispiele für Antwortverzerrungen durch Gegenüberstellung extremer und gemäßigter Antwortka-

ategorien beschreiben Kreutz & Titscher (1974) und Payne (1951). Besondere Vorsicht ist geboten, wenn extreme Kategorien in ihrer Formulierung Existenzaussagen („Es gibt . . .“) oder Allaussagen („Alle . . .“, „immer . . .“, „nie . . .“) nahekommen (Payne 1951).

Mit der häufig beobachteten Tendenz der Befragten, extreme Antwortkategorien zu meiden (d.h. sie eher zu wählen, wenn noch extremere vorgegeben sind), hängt es auch zusammen, daß nachträgliche Kombinationen von Antwortkategorien (z.B. „sehr dafür“ und „dafür“) fast immer andere Ergebnisse liefern als Befragungen, in denen von vornherein eine zusammengefaßte Kategorie („sehr dafür oder dafür“) vorgegeben war. Welche Ergebnisse die „richtigen“ sind, ist in der Regel natürlich nicht entscheidbar (Payne 1951).

Sind bei Wissensfragen die richtigen Antworten in den Vorgaben enthalten, so ergeben sich selbst wenn sie nur dem Interviewer für Zwecke der Feldverschlüsselung vorliegen, größere Häufigkeiten richtiger Antworten als im Falle offener Fragen (Noelle-Neumann 1970).

Da nicht davon auszugehen ist, daß jede Vp zu jeder Frage eine Antwort geben kann, ist es einerseits zur Vermeidung artifizieller (zufälliger) Wahl von Antworten erforderlich, Restkategorien vorzusehen, andererseits ermöglichen solche Kategorien den Vpn ein „Ausweichen“ (und provozieren es u.U. sogar), so daß in der Praxis häufig darauf verzichtet wird (Rugg & Cantril 1972, Kirschhofer-Bozenhardt & Kaplitza 1975).

Tatsächlich müßten zur Abdeckung aller denkbaren Fälle sogar mehrere Ausweichkategorien vorgesehen werden (Karmasin & Karmasin 1977): Galtung (1973) unterscheidet zwischen kognitiven („weiß nicht“) und evaluativen („interessiert mich nicht“) Nicht-Antworten, dazu müßte man noch berücksichtigen, daß „etwas anderes“ oder „mehreres“ für die Vp richtig sein oder sie die Frage nicht verstanden haben kann („nicht verstanden“). Die Hinzufügung von Nicht-Antwortkategorien zu einem Satz von Antwortvorgaben führt u.U. zu beträchtlichen Wahlhäufigkeiten für diese und entsprechenden Veränderungen für die anderen Kategorien, wobei nur im Einzelfall geklärt werden kann, ob dadurch wahre Varianz (Bequemlichkeitshypothese) oder Fehlervarianz (Hypothese des Zufallscharakters erzwungener Antworten) von den inhaltlichen Kategorien abgezogen wird.

Zahlreiche Untersuchungen zeigen auf, daß die Reihenfolge der Antwortalternativen im Fragentext in einer Listenvorgabe einen Einfluß auf die Wahlhäufigkeiten ausübt. Dabei variieren die Angaben vor allem über das genaue Ausmaß solcher „Positionseffekte“ von Untersucher zu Untersucher beträchtlich. Wegen der zahlreichen Interaktionen mit inhaltlichen und formalen Aspekten der Fragen ist auch nicht damit zu rechnen, daß allgemein gültige Aussagen möglich sind (Kreutz & Titscher 1974). Payne (1951) spricht von Tendenzen

dahingehend, daß bei mündlichen qualitativen Vorgaben die zuletzt genannten, bei schriftlichen qualitativen Kategorien diejenigen in Extrempositionen (erste und letzte Stelle) relativ inhaltsunabhängig bevorzugt werden. Bei quantitativen Vorgaben gibt es relativ unabhängig von ihrer Höhe eine Neigung, solche in der Nähe des Mittelwertes zu wählen. In diese Richtung gehen auch die Befunde von Stroschein (1964), Belson (1966) und Ring (1974).

Zur Vermeidung systematischer Auswirkungen solcher Positionseffekte hat es sich eingebürgert, sogenannte gegabelte Befragungen (split-ballot, vgl. 3.4) durchzuführen und dabei die Reihenfolge von Vorgaben zu variieren. Wegen der Benachteiligung der Mittelpositionen ist dabei ein einfaches ‚Umdrehen‘ meist nicht ausreichend (‚verfeinerte‘ Techniken beschreibt Ring 1974). Dieses Verfahren findet allerdings dort seine Grenzen, wo die Frage unnatürlich zu wirken beginnt (Payne 1951) und befürchtet werden muß, daß dadurch zusätzliche Störeinflüsse wirksam werden (z.B. ‚Gehen Sie heute abend nicht ins Theater oder gehen Sie ins Theater?‘). Da Positionseffekte sich dann nicht zeigen, wenn über jede Vorgabe durch Einzelantwort entschieden werden muß (Stroschein 1965), bietet sich dieser Fragentyp für Fälle an, in denen die Variation der Vorgabenreihenfolge nicht möglich ist. Neben der Ausgewogenheit der Kategorien hat auch die Ausgewogenheit der mit den jeweiligen Antworten verbundenen Folgefragen einen erheblichen Einfluß auf die Wahlhäufigkeit (Noelle-Neumann 1970). Vor allem scheinen Befragte (im mündlichen Interview evtl. auch Interviewer) rasch die Vermeidung bestimmter Antworten (z.B. ‚Ja‘) zu lernen, wenn diese regelmäßig mit einer höheren Zahl von Folgefragen verknüpft sind (Cannell & Kahn 1968).

3.1.4 *Verzerrte Fragen*

Verzerrte Fragen sind nach Friedrichs (1973, 198) solche, „... die allein durch ihre Formulierung die Verteilung der Antworten in einer bestimmten Form beeinflussen ...“. Solche Verzerrungen können einmal durch inadäquate Antwortkategorien (vgl. dazu 3.1.3) entstehen, daneben gibt es weitere Faktoren, für die verzerrende Wirkungen auf Antworten aufgezeigt worden sind.

- a. *Unterstellungen* (Implikationen) führen, soweit sie zu Unrecht bestehen, häufig nicht dazu, daß sie von den Vpn zurückgewiesen werden, sondern verzerren die Antworten. Das gilt für die unterstellte Vertrautheit mit Sachverhalten und Begriffen (Payne 1951, vgl. dazu speziell 3.1.5) ebenso wie für unterstellte Voraussetzungen (‚Welcher Teil Ihrer Arbeit stört Sie am meisten?‘ - wer sagt, daß einer stört?) und unterstellte Konsequenzen eines zu erfragenden Sachverhaltes (‚Welcher Teil Ihrer Arbeit stört Sie am meisten, d.h. welchen schieben Sie am längsten auf?‘). Letztere entstehen häufig versehentlich beim Versuch einer Konkretisierung des Befragungsgegenstandes (Payne 1951).

Aber auch wenn Unterstellungen nicht verzerrend wirken, sondern von Befragten erkannt und zurückgewiesen werden (z.B. ‚Wieviele Zigaretten rauchen Sie pro Tag?‘), stellen sie einen befragungstechnischen Fehler dar. In bestimmten Fällen, z.B. um bestimmten peinlichen oder sozial unerwünschten Verhaltensweisen den Charakter des Selbstverständlichen zu verleihen (vgl. Kinsey et al. 1970), können Unterstellungen als bewußt eingesetztes methodisches Hilfsmittel gerechtfertigt sein. Auf die ansonsten erforderliche Vorschaltung einer Filterfrage wird dann zurecht verzichtet. Phillips (1966) erwähnt als Beispiel die Frage nach finanziellen Belastungen aus Ratenkäufen. Statt der üblichen Fragenfolge ‚Haben Sie regelmäßige Zahlungsverpflichtungen aus Ratenkäufen?‘ (= Filter), ‚wenn ja: Wie hoch sind diese pro Monat?‘, wäre es hier zweckmäßig, zur Verminderung des Einflusses der sozialen Erwünschtheit mit einer Unterstellung zu arbeiten und zu fragen: ‚Wie hoch sind Ihre Zahlungsverpflichtungen aus Ratenkäufen pro Monat?‘.

- b. Die Verknüpfung bestimmter Befragungsgegenstände mit *Persönlichkeiten, wichtigen Ereignissen* o.ä. hat - wie in verschiedenen Untersuchungen nachgewiesen - einen erheblichen Einfluß auf die Antworten (vgl. auch 2.3). So berichten Roslow et al. (1940) und Rugg & Cantril (1972) über Auswirkungen, die die Erwähnung des ‚Präsidenten‘ bzw. des ‚Kongresses‘ in Fragen zu aktuellen politischen Problemen zeigten. Dabei sind diese Auswirkungen allerdings spezifisch, d.h. sie zeigten sich nicht durchgängig bei beliebigen Befragungsgegenständen (Rugg & Cantril 1972). Mit erheblich verändertem Antwortverhalten wäre z.B. auch zu rechnen, würde man eine Frage über Sicherheitsmaßnahmen von Fluggesellschaften an einer aktuellen Flugzeugentführung ‚festmachen‘ (Karmasin & Karmasin 1977).
- c. *Affektiv getönte Begriffe* vermögen Antwortverteilungen deutlich zu beeinflussen. So berichten Rugg (1941) und Rugg & Cantril (1972) niedrigere Zustimmungsraten für ‚verbieten‘ vs. ‚nicht erlauben‘ (im Zusammenhang mit ‚öffentlichen Reden gegen die demokratische Ordnung‘) und für ‚den Krieg erklären‘ vs. ‚in den Krieg eintreten‘ (im Zusammenhang mit dem Eintritt der USA in den Zweiten Weltkrieg). Besonders auffallend ist die Wirkung des Begriffes ‚Veränderung‘: Die Frage nach einer *Ergänzung* der Verfassung um eine bestimmte Vorschrift (‚hinzufügen‘) erhielt 36% Zustimmungen und 50% Ablehnungen, die nach einer (inhaltlich identischen) Verfassungsänderung aber nur 26% Zustimmungen und 65% Ablehnungen (Rugg & Cantril 1972, 106). Dieser und ähnliche Befunde veranlaßten Payne (1951, 183) prinzipiell alle Fragen, die explizit entweder auf den ‚Status-Quo‘ oder auf ‚Veränderungen‘ (oder auf beides) hinweisen, schon allein deshalb für verzerrt zu halten.

Daneben kann natürlich fast jeder Begriff in einem bestimmten gegebenen Befragungszusammenhang affektiv getönt sein. Falls entsprechende Befürchtungen begründet sind, empfiehlt sich im Rahmen des Pretests eine entsprechende Untersuchung z.B. unter Verwendung der Methode des se-

mentischen Differentials oder durch Bestimmung von Assoziationen (vgl. Kreutz & Tischer 1974).

- d. Verzerrt kann eine Frage schließlich auch dadurch sein, daß sie durch ihren Aufbau bestimmte Antworten deutlich begünstigt („Lehnen Sie es ab . . .“, „Sie sind doch auch der Meinung . . .“ etc.).

Litwak (1956) hat darauf hingewiesen, daß die Verzerrtheit einer Frage nicht nur von Merkmalen dieser Frage, sondern auch von ihrem Verwendungszweck abhängt: Was im Rahmen einer demoskopischen Befragung eine „unerlaubte“ verzerrte Frage wäre, kann innerhalb einer Einstellungsskala ein zulässiges, sogar erforderliches extremes Item sein. Auch in anderen Zusammenhängen können verzerrte Fragen (bewußt eingesetzt und die Ergebnisse entsprechend interpretiert) zu bemerkenswerteren und gültigeren Erkenntnissen führen, als ausgewogene Fragen dies tun würden (vgl. a.a.O. und Anger 1969, Kreutz & Tischer 1974). So ist es denkbar, daß bezüglich bestimmter Sachverhalte (z.B. Kernkraft) weniger die evtl. stark von Medien beeinflussten und u.U. wenig stabilen Reaktionen des „Durchschnittsbürgers“ auf ausgewogene Fragen und eher die Sichtweisen eines durch verzerrte Fragen herausgefilterten „harten Kerns“ von Gegnern und Befürwortern interessieren.

Außerdem ist es möglich, verzerrte Fragen bzw. Fragebogen mit verzerrten Fragen nicht zum Zwecke der Informationsgewinnung, sondern mit dem Ziel der Beeinflussung im Sinne einer Einstellungsänderung einzusetzen. Über frühe derartige Versuche (Beeinflussung interventionistischer vs. isolationistischer und gewerkschaftsfreundlicher vs. gewerkschaftsfeindlicher Einstellungen) von Roper berichten Rugg & Cantril (1972). Einstellungsänderungen ließen sich vor allem bezüglich solcher Sachverhalte erzielen, denen gegenüber die Vpn verhältnismäßig unsicher waren. Dillehay & Jernigan (1970) konnten durch einen verzerrten Fragebogen zur Behandlung von Straftätern nur Einstellungsänderungen in Richtung auf mildere, nicht aber solche in Richtung auf härtere Bestrafung erzielen. Selbstverständlich darf dabei die Verzerrung der Fragen nicht soweit gehen, daß sie von den Vpn als Beeinflussungsversuch erkannt wird (sonst wäre durchaus auch mit Bumerangeffekten zu rechnen).

Diese als Beispiele für den absichtlichen Einsatz verzerrter Fragen erwähnten Untersuchungen stellen erneut die Beeinflußbarkeit des Antwortverhaltens der Vpn durch konstruktive Merkmale der Frage unter Beweis, machen andererseits aber auch deutlich, daß dieses Problem vor allem bei unsicheren Beurteilungsgrundlagen, Einstellungen oder Meinungen besteht und daß Merkmale der Fragen an Bedeutung verlieren, wenn die zu erfragenden Inhalte deutlich ausgeprägt, stabil bzw. intensiv sind (vgl. dazu auch 3.4).

3.1.5 Uninformiertheit, Meinungslosigkeit und Urteilsausgewogenheit

Die Tatsache, daß Vpn eine Frage beantworten, kann nicht als Hinweis darauf interpretiert werden, daß sie über den erfragten Sachverhalt informiert sind oder eine Meinung dazu haben. So berichtet Payne (1951, 156) über Nonsense-Fragen (z.B. Beurteilung eines nicht existenten Gesetzentwurfs), die gleichwohl von erheblichen Anteilen der Vpn ‚beantwortet‘ wurden. Ein besonders eindrucksvolles Beispiel beschreibt Eysenck (1956, 156). Bei einer Umfrage in Großbritannien kurz nach dem Ende des Zweiten Weltkriegs wurde die Frage gestellt, ob man ‚König Georg von Griechenland‘ wieder in sein Land zurückkehren lassen sollte. 60% der Befragten bejahten diese Frage. In einer etwa gleichzeitig durchgeführten anderen Befragung gab jedoch nur ein kleiner Bruchteil der Befragten an, schon einmal etwas von König Georg von Griechenland gehört zu haben. Offenbar neigen also Befragte dazu, ihre Uninformiertheit nicht zu offenbaren (und Fragebogenkonstrukteure geben ihnen häufig auch gar keine Möglichkeit, dies zu tun). Einer Frage z.B. nach der Beurteilung eines Sachverhaltes muß deshalb entweder eine Filterfrage nach der Informiertheit, eine Unterweisungsfrage (vgl. 2.1) oder eine Erklärung vorangehen, wobei nach Noelle-Neumann (1974) die Unterweisungsfrage (‚... wissen Sie davon?‘) im Vergleich zur Erklärung der effizientere Weg ist, da sie eine aktive Auseinandersetzung mit der Information (eine Antwort) erfordert.

In jedem Fall ist dabei auf das Bedürfnis der Vpn, informiert zu erscheinen, Rücksicht zu nehmen, d.h. ein Bloßstellen der uninformierten Vpn muß vermieden werden.

Dies kann z.B. wiederum dadurch geschehen, daß in einer Filterfrage der Uninformiertheit der Charakter des Selbstverständlichen verliehen wird (vgl. Maccoby & Maccoby 1972). Statt ‚Wissen Sie, welche Länder Mitglieder der EG sind oder wissen Sie das nicht?‘ wäre eine Formulierung ‚Wissen Sie vielleicht, welche Länder Mitglieder der EG sind?‘ vorzuziehen (Karmasin & Karmasin 1977, vgl. auch Phillips 1966). Eine Erklärung im Rahmen einer Frage sollte aus den selben Gründen nicht belehrend wirken (‚Unter EG versteht man die Europäische Gemeinschaft . . .‘), sondern entweder die Vermutung ihrer Entbehrlichkeit zum Ausdruck bringen (‚Wie Sie vermutlich wissen, ist die EG . . .‘) oder aber als Präzisierung des Befragungsgegenstandes in Erscheinung treten (‚Was halten Sie von der Europäischen Gemeinschaft, also der EG?‘).

Neben der gegebenen, aber durch inadäquate Fragenkonstruktion unerkannt bleibenden Uninformiertheit von Vpn ist die verbreitete Unterstellung, Befragte würden zu ausnahmslos allen Befragungsgegenständen eine Meinung haben, eine häufige Ursache für inadäquate Antworten (Kreutz & Titscher 1974). Auch hier muß die Konstruktion der Frage es der Vp in geeigneter

Weise ermöglichen, das Nichtvorhandensein einer Meinung zum Ausdruck zu bringen (zur Vorgabe entsprechender Antwortkategorien vgl. 3.1.3).

Empirische Belege sprechen außerdem dafür, daß Vpn deutlich negative Urteile über Befragungsgegenstände scheuen (vgl. Roslow et al. 1940, Rugg 1941, Phillips 1966, Kirschhofer-Bozenhardt & Kaplitza 1975), auch wenn sich das nicht durchgängig demonstrieren läßt (z.B. die Formulierung ‚X ist besser als Y‘ nicht generell der logisch äquivalenten ‚Y ist schlechter als X‘ vorgezogen wird, vgl. Adams 1956), und eher zu ausgewogenen Beurteilungen neigen. Es ist deshalb erforderlich, Befragte nicht ausschließlich zur Kritik an Befragungsgegenständen zu zwingen, sondern Gelegenheit zur Hervorhebung positiver Aspekte zu geben, auch wenn solche inhaltlich gar nicht interessieren sollten (Phillips 1966; Noelle-Neumann 1974 spricht dann von ‚Wegwerf-Fragen‘).

3.1.6 Antworttendenzen und vorschnelle Antworten

Bei Sachverhalten, bei denen eine wahrheitsgemäße Antwort nach Meinung der Vp gleichzeitig eine sozial unerwünschte Antwort wäre, muß durch geeignete Fragenkonstruktion der Vp z.B. die Möglichkeit gegeben werden, sich für die Antwort zu ‚entschuldigen‘. Statt zu fragen ‚Besitzen Sie ein Auto oder besitzen sie ein solches nicht?‘ und damit der Vp evtl. das Eingeständnis ihrer ‚Armut‘ abzuverlangen, könnte man an folgende Konzeption denken: ‚Besitzen Sie ein Auto oder ist das für Sie im Augenblick nicht möglich oder wünschenswert?‘ (vgl. dazu auch Phillips 1966).

Um response sets wie der Bejahungstendenz entgegenzuwirken und außerdem die Vpn zu sorgfältiger Beantwortung der Fragen zu veranlassen, wird meist empfohlen, Items teilweise positiv (‚Sind Sie an Sport interessiert?‘), teilweise negativ (‚Sind Sie an Sport uninteressiert?‘) zu formulieren. Allerdings ist eine solche Vorgehensweise nicht unproblematisch. Zum einen ergibt sich durch die Antwort ‚Nein‘ auf ein negativ formuliertes Item die Situation der doppelten Verneinung, die stets als Fehlerquelle anzusehen ist (vgl. 3.2.2). Zum andern haben Terborg & Peters (1974) gezeigt, daß die Veränderung der Formulierungsrichtung für viele Items signifikante Auswirkungen auf die Häufigkeit der Wahl von Antworten hat, d.h. die Antwort ‚Ja‘ auf ein positiv formuliertes Item nur logisch äquivalent der Antwort ‚Nein‘ auf ein negativ formuliertes (und umgekehrt) ist. Diese Beantwortungsunterschiede konnten zudem nicht ausschließlich der unterschiedlichen Wirksamkeit von Antworttendenzen (‚Ja-Tendenz oder Nein-Tendenz‘) angelastet werden, da je nach Item Ja-Antworten bei positiver Formulierung häufiger, z.T. aber auch seltener auftraten als Nein-Antworten bei negativer Formulierung (und umgekehrt). Karmasin & Karmasin (1977) schlagen deshalb vor, auf Ja-Nein-Fragen möglichst

zu verzichten und stattdessen die Alternativen explizit zu formulieren, wobei deren Reihenfolge im Rahmen einer gegabelten Befragung (vgl. 3.4) variiert werden kann (die o.a. Frage würde dann lauten: „Interessieren Sie sich für Sport oder sind Sie an Sport uninteressiert?“). Prinzipiell ähnliche Überlegungen für den Fall von Persönlichkeitsfragebogen finden sich bei Ehlers (1973) und Keil (1973).

Einer zweiten Tendenz im Beantwortungsverhalten sollte ebenfalls schon durch die Konstruktion der Frage entgegengewirkt werden: der Neigung zu impliziter oder expliziter Formulierung der Antwort, bevor die gesamte Frage von der Vp zur Kenntnis genommen wurde (Tendenz zu vorschneller Antwort). Das kann z.B. dadurch erfolgen, daß die eigentliche Frage erst ganz zum Schluß, d.h. nach der genauen Definition des Gegenstandes, der Explikation des Bezugsrahmens etc. verbalisiert wird (Payne 1951). Dem zugegebenermaßen konstruierten Aufbau einer Frage

- „Würden Sie sagen, der Preis für Benzin ist zu hoch (*), gerade richtig oder zu niedrig (*), wenn Sie ihn mit Preisen anderer Dinge vergleichen?“, der mindestens an den mit (*) bezeichneten Stellen vorschnelle Antworten ermöglicht, wäre ein Aufbau wie etwa der folgende vorzuziehen:
- „Verglichen mit den Preisen anderer Dinge: Würden Sie sagen, der Preis für Benzin ist zu hoch, gerade richtig oder zu niedrig?“.

3.2 Sprachliche Formulierung der Frage

3.2.1 Kriterien für die sprachliche Formulierung

Die sprachliche Formulierung einer Frage erfolgt einerseits mit dem Ziel, den Befragten zu einer Antwort zu motivieren, andererseits muß sie erreichen, daß die Frage von der Vp richtig verstanden wird (Anger 1969). Übereinstimmend wird die sprachliche Formulierung einer Frage als ein Problem der Optimierung unter dem Kriterium der Bedeutungsäquivalenz für alle Befragten angesehen. Da die Bedeutung eines Begriffes außer von seiner Denotation (definierbarem Inhalt) und den interindividuell unterschiedlichen Konnotationen (dem Bedeutungshof, der den Ort in einem semantischen Raum bestimmt, vgl. Osgood et al. 1957) auch noch von gruppenspezifischen Bedeutungsanteilen (Altersgruppen, Schichten, regionalen Gruppierungen; Karmasin & Karmasin 1977) und der Verwendung in unterschiedlichen Lebensbereichen (Arbeitswelt, Privatleben etc.; vgl. Scheuch 1973) geprägt wird, ist es grundsätzlich ausgeschlossen, das absolute Optimum der Bedeutungsäquivalenz (inhaltliche Standardisierung) durch identische sprachliche Formulierung für alle Vpn (formale Standardisierung im Sinne von Stroschein 1965) zu erreichen. Vielmehr würde dazu eine für jede Vp unterschiedliche sprachliche Formulierung erforderlich sein. Ausgehend von der Erfahrung, daß in der alltäglichen Kommuni-

kation wechselseitiges Verstehen mehr oder weniger möglich ist, wird im freien mündlichen Interview die Aufgabe, eine inhaltliche Standardisierung in dem genannten Sinne zu erreichen, der Intuition des Interviewers übertragen. Der dagegen erhobene Einwand, daß hierbei (von einzelnen hochqualifizierten und ‚begabten‘ Interviewern vielleicht abgesehen) zu der Störvariable ‚Verständnisunterschiede durch die Vpn‘ nur noch weitere hinzugefügt würden (wie etwa die Ausdrucksfähigkeit des Interviewers und seine Vorstellung von dem, was die Vp wie zu verstehen habe: Wottawa 1980), läßt sich jedoch kaum entkräften.

Nach dem Prinzip der „maximalen Übelminimierung“ (Wottawa 1980, 209) werden bei der formalen Standardisierung (Identität der Fragen auf verbaler Ebene) die interindividuellen sprachlichen Unterschiede vernachlässigt. Ziel bleibt auch hier die Bedeutungsäquivalenz, nur liegt diesem Vorgehen die Annahme zugrunde, daß diese durch identische sprachliche Formulierungen besser als durch unkontrollierte freie Formulierungen zu erreichen sei (Mayntz et al. 1971, Wottawa 1980). Dies wird in der Regel auch für Instruktionen im Rahmen von Leistungstests oder von psychologischen Experimenten angenommen. Da es andererseits offensichtlich unsinnig ist, eine Frage beantworten zu lassen, die nicht verstanden wurde, ergibt sich fast zwangsläufig die Forderung, sich bei der sprachlichen Formulierung an der untersten Grenze der Zielgruppe zu orientieren (Payne 1951). Allerdings genügen solche Formulierungen mindestens für sprachlich differenziertere Vpn nicht mehr dem Kriterium der Motivierung von Antworten (Erdos 1970), da durch Übersimplifizierungen Zweifel an der Seriosität der Befragung ausgelöst werden können (Kreutz & Titscher 1974). Deshalb wird als Kompromiß heute eher eine Orientierung an der Alltagssprache (Umgangssprache) des durchschnittlichen Mitgliedes der Zielpopulation vorgeschlagen (Karmasin & Karmasin 1977).

Teilweise wird versucht, gruppenspezifische Bedeutungsunterschiede durch unterschiedliche sprachliche Formulierungen zu berücksichtigen. Dies gilt vor allem für regionale Unterschiede. Noelle-Neumann (1963, 1974) etwa schlägt vor, den Interviewer durch eine sogenannte informelle Ermittlung (eine Frage ohne festgelegten Wortlaut, z.B. nach der Gebräuchlichkeit der Bezeichnungen ‚Samstag‘ oder ‚Sonntag‘) die Zugehörigkeit des Befragten zu einem bestimmten Sprachraum feststellen zu lassen und in Abhängigkeit davon zu verschiedenen formulierten Fragen zu verzweigen. Im übrigen muß aber in der Regel (mit Karmasin & Karmasin 1977, 176) festgestellt werden: „... über die Bedeutungsverschiebungen von einzelnen Begriffen bzw. über die jeweils relevanten wörtlichen Bezeichnungen von Sachverhalten in den üblichen Sprachrepertoires von Jugendlichen gegenüber Erwachsenen, Männern gegenüber Frauen, Unterschicht gegenüber Oberschicht ist jedoch im Augenblick aus dem deutschen Sprachraum noch zu wenig bekannt, so daß auch hier nur der Ausweg bleibt, wörtliche Äquivalenz zu wahren und bei allen Begriffen und

Formulierungen, bei denen wechselnde Bezugsrahmen vermutet werden können, den Bezugsrahmen mit anzugeben, unter dem der Forscher den Begriff einzuordnen wünscht“ (vgl. auch Cannell & Kahn 1968). Mangels gesicherten Wissens über sprachliche Unterschiede ist häufig nicht ohne weiteres zu entscheiden, ob Beantwortungsunterschiede z.B. zwischen Angehörigen verschiedener Schichten auf unterschiedliches Fragenverständnis oder auf Unterschiede des erfragten Sachverhaltes zurückgehen (Kreutz & Titscher 1974). Mindestens empfiehlt es sich in solchen Fällen, einen Sachverhalt durch mehrere Fragen unterschiedlicher Formulierung zu erfassen bzw. Verständnis-Kontrollfragen einzubauen.

3.2.2 Anforderungen an die sprachliche Formulierung

Wegen der starken Kontextabhängigkeit und der Vielfalt der Interaktionen mit inhaltlichen Aspekten scheint es von vornherein verfehlt, nach besonders geeigneten Standardformulierungen zu suchen (Raab 1974). Stattdessen wird man die Argumente für oder gegen bestimmte Vorgehensweisen im Einzelfall gegeneinander abwägen müssen. Folgt man dem Prinzip, sich bei der sprachlichen Formulierung an der Umgangssprache des Durchschnitts der Zielpopulation zu orientieren (Karmasin & Karmasin 1977, vgl. auch 3.2.1), so muß man sich zunächst die hauptsächlichen Kennzeichen dieser Sprache vergegenwärtigen (eine entsprechende Zusammenstellung unter Berücksichtigung der Ergebnisse von Lesbarkeitsuntersuchungen findet sich - allerdings für den anglo-amerikanischen Sprachbereich - z.B. bei Wright & Barnard 1975).

Dazu gehört - mindestens im Falle von Wohnbevölkerungen als Zielgruppen - die Verwendung kurzer Wörter. Payne (1951) berichtet, daß Fragen, die unter dem Kriterium geringer Beeinflußbarkeit der Antworthäufigkeiten durch Variation der Vorgabenreihenfolge als ‚klar‘ klassifiziert worden waren, zu ca. 8%, ‚unklare‘ Fragen dagegen zu 12,5% zwei- oder mehrsilbige Wörter enthielten. Bei klaren Fragen waren 30%, bei unklaren 40% aller Silben Vor- oder Nachsilben. Sodann ist für die Alltagssprache der Gebrauch solcher Wörter charakteristisch, die in der Sprache häufig vorkommen. Dementsprechend schließt sich auch Friedrichs (1973) im Zusammenhang mit der Fragenformulierung der Empfehlung einer Beschränkung auf die 1000 gebräuchlichsten Wörter (mit Vorbehalten) an. Das Kriterium der Worthäufigkeiten ist indessen recht oberflächlich, da eigentlich relevanter die Gebräuchlichkeit des Wortes in einem gegebenen Zusammenhang ist. Fremdwörter und Abstrakta jedenfalls sollten soweit als möglich vermieden werden.

Weniger eine Frage der Orientierung an der Alltagssprache als eine Notwendigkeit im Hinblick auf die Approximation der Bedeutungsäquivalenz für alle Befragten ist die Notwendigkeit einer Beschränkung auf klare Begriffe, das sind solche, deren denotative Bedeutungen prägnant und die arm an konno-

tativen Bedeutungen sind (Rohrmann 1978), was gegebenenfalls durch Analysen z.B. unter Verwendung des semantischen Differentials nachzuprüfen wäre (Friedrichs 1973). Im Interesse dieser Klarheit empfiehlt es sich auch nicht, Synonyme im Wechsel zu verwenden (Anger 1969).

Zwar herrschen in der Alltagssprache personalisierte Formulierungen vor (Karmasin & Karmasin 1977), doch muß hier die Entscheidung in Abhängigkeit von der Fragestellung erfolgen, da mit spezifischen Einflüssen der Personalisierung einer Frage auf die Antworten zu rechnen ist (vgl. 3.1.3 und 3.3).

Die grammatikalische Satzkonstruktion der Alltagssprache ist durch geringe ‚Satztiefe‘ (geringen Komplexitätsgrad der syntaktischen Struktur, vgl. Karmasin & Karmasin 1977) ausgezeichnet. Ob man deshalb mit Kreutz & Titscher (1974) für Fragen eine Beschränkung auf Hauptsätze fordern muß, ist zweifelhaft. Immerhin sollten

- ungewöhnliche Tempora,
- komplizierte Nebensatzkonstruktionen (Schachtelsätze),
- adverbiale Konstruktionen und
- passivische Formulierungen

möglichst vermieden werden (Wright & Barnard 1975, Karmasin & Karmasin 1977). Dasselbe gilt für doppelte Verneinungen, die entweder im Fragentext selbst liegen (‚Sind Sie dagegen, daß der 17. Juni als Feiertag abgeschafft wird oder sind Sie nicht dagegen?‘) oder durch eine verneinende Antwort auf eine negativ formulierte Frage entstehen können (‚Soll der 17. Juni in Zukunft kein Feiertag mehr sein? ja/nein‘, vgl. auch 3.1.6). In vielen Fällen genügt es allerdings nicht, eine grammatikalisch richtige Fragenkonstruktion zu verwenden, zusätzlich muß auch sichergestellt sein, daß der Bezug der Antwort auf die Frage unmittelbar evident ist. So ist nach Payne (1951, 69) bei einer Frage des Typs ‚Ist Ihr Gesundheitszustand heute besser oder schlechter als vor einem Jahr?‘ trotz grammatikalischer Eindeutigkeit vielen Vpn nicht klar, ob sich die Antwort ‚besser‘ auf ‚heute‘ oder auf das vergangene Jahr bezieht. In solchen Fällen ist es unabdingbar, daß die Alternativen explizit formuliert werden (‚Ist Ihr Gesundheitszustand heute besser oder war er vor einem Jahr besser?‘).

Vielfach findet sich in der Literatur die Empfehlung, Fragen möglichst kurz zu fassen (z.B. Holm 1974b, Kreutz & Titscher 1974, Wright & Barnard 1975, Karmasin & Karmasin 1977). Oppenheim (1966) empfiehlt 20 Wörter als Obergrenze, Payne (1951) berichtet für die nach seinem Kriterium ‚klaren‘ Fragen (s.o.) eine durchschnittliche Länge von 22, für ‚unklare‘ eine von 31 Wörtern. Schneider-Düker & Schneider (1977) fanden bei ihren Versuchen zur freien Reproduktion von Fragebogenitems Korrelationen von 0,54 bzw. 0,71 zwischen Anzahl von Intrusionen (Umformungen bzw. Einfügungen von Wörtern) und Itemlänge (Wortanzahl bzw. Silbenanzahl).

Für offene Fragen und Listenfragen mit Mehrfachnennungen in mündlichen standardisierten Interviews widersprechen Cannell et al. (1977) der Forderung nach möglichst kurzen Items. Sie stellen der dieser Forderung zugrundeliegenden Hypothese von der ‚Verwirrung‘ der Vp durch eine lange Frage eine Hypothese der ‚Vorbildwirkung‘ des Interviewers gegenüber und vermuten, daß die Vp ihr Engagement und ihre Ausführlichkeit bei der Beantwortung an den cues orientiert, die sie aus dem Verhalten des Interviewers entnimmt. Beim Vergleich der Antworten auf Items in Kurzform (durchschnittlich 14 Wörter) und Langform (= Kurzform + Redundanz, durchschnittlich 38 Wörter) zeigte sich zwar kein Unterschied in den Antwortlängen, dafür enthielten freie Antworten und Antworten in Listenfragen im Falle der Langform mehr Information. Außerdem waren auch die Antworten auf kurze Fragen informationshaltiger, wenn der Fragebogen teils kurze, teils lange Fragen enthielt. Soweit es in den Fragen um Reproduktion von Gedächtnisinhalten geht, ist auch zu bedenken, daß eine lange Frage der Vp mehr Zeit läßt und u.U. relevante cues mehrfach wiederholt darbietet (Cannell et al. 1977). Allerdings wirkt sich die Fragenlänge möglicherweise nicht auf alle Vpn gleichmäßig aus. Koomen & Dijkstra (1975) z.B. fanden (anders als Cannell et al. 1977) einen Anstieg der Antwortlängen in Abhängigkeit von der Länge der Fragen, allerdings nur für solche Vpn, die bei kurzen Fragen zu ausgesprochen kurzen Antworten neigten (vgl. auch Sudman & Bradburn 1974).

Auch zur Fragenlänge lassen sich keine unbeschränkt gültigen Aussagen machen. Sie ist unter Berücksichtigung von Frageninhalt, Untersuchungsziel und Verständlichkeit im Einzelfall zu optimieren.

3.3 Spezielle Gesichtspunkte der Formulierung von Items für diagnostische Fragebogen

Für diagnostische Fragebogen, die auf der Grundlage eines streng empirischen Validitätskonzeptes erstellt wurden, ist die Kriteriumskorrelation der Prüfstein für die adäquate Formulierung eines Items. Theoretisch ist es hier sogar zulässig, vollkommen unverständliche Items zu konstruieren, solange nur die ‚Art des Unverständnisses‘ (z.B. der dann die Antwort determinierende response set, vgl. 1.2.1 und 1.2.2) gültige Vorhersagen ermöglicht.

Wird von inhaltlichen Überlegungen ausgegangen, so muß sich letztlich in Itemanalysen (entweder im Rahmen eines klassischen oder eines probabilistischen Meßmodelles) die inhaltliche und formale Brauchbarkeit der Items erweisen. Die Wahrscheinlichkeit dafür, daß dies gelingt, dürfte bei Berücksichtigung der in diesem Kapitel aufgezeigten Zusammenhänge erhöht sein.

Im Hinblick auf die Motivation der Probanden müssen mehr noch als bei demoskopischen Fragebogen Überlegungen zur face-validity der Items ange-

stellt werden. Auch bei den demoskopischen Fragebogen spielt dieser Aspekt eine Rolle: Nicht jedem Untersucher wird von den Befragten jede Frage ‚zugestanden‘ (vgl. Richter 1969). Für rein empirische Fragebogenkonstruktionen ist face-validity im Interesse der Undurchschaubarkeit der Items geradezu unerwünscht und oft auch tatsächlich nicht gegeben, was solche Verfahren bei Betroffenen und in der Öffentlichkeit häufig in Mißkredit bringt. Empfehlungen für die Formulierung von Items in diagnostischen Fragebogen und von Aufgaben in Leistungstests finden sich in der entsprechenden Literatur, z.B. bei Lienert (1969, 62ff) und Wottawa (1980, 212ff).

3.4 Die Kontrolle von Formulierungseinflüssen

Während im Zusammenhang mit Einstellungsskalen (vgl. z.B. Suchman & Guttman 1947) und diagnostischen Verfahren (z.B. soweit sie auf der Basis eines probabilistischen Meßmodelles, vgl. Fischer 1974, Wottawa 1980, konstruiert sind) auch grundlegend andere Strategien verfolgt werden, versucht man im Fall demoskopischer Fragebogen die vielfältigen Einflüsse der Fragenformulierung auf Antworten durch ‚Mittelungsprozeduren‘ zu eliminieren. Schon in den 30er Jahren (vgl. z.B. Roslow et al. 1940, Rugg 1941) wurde damit begonnen, innerhalb einer Befragung unterschiedliche Fragenformulierungen zu verwenden (gegabelte Befragung, split-ballot-verfahren) und als Ergebnis einen Mittelwert aus den in der Regel differierenden Antworten zu verwenden. Heute ist diese Vorgehensweise weithin üblich (vgl. z.B. Payne 1951, Stroschein 1965, Noelle-Neumann 1963, 1970, Rugg & Cantril 1972). Karmasin & Karmasin (1977) referieren eine Untersuchung, in der mit 12 verschiedenen Varianten des Fragebogens (variieren Reihenfolgen von Antwortkategorien) gearbeitet worden ist. Zweifel am Sinn dieses Verfahrens äußert allerdings bereits Noelle-Neumann (1970). Es fragt sich, was ein auf diese Weise zustande gekommenes ‚mittleres‘ Ergebnis eigentlich bedeutet. Es wäre sinnvoll nur interpretierbar, handelte es sich bei den Formulierungseffekten um ‚Zufallsfehler‘ mit einem Erwartungswert von Null, tatsächlich aber muß angenommen werden, daß mit unterschiedlich formulierten Fragen Unterschiedliches gemessen wird (Raab 1974), sonst dürften die Antwortunterschiede in Abhängigkeit von der Formulierung nicht (wie in diesem Kapitel oft berichtet) signifikant bzw. konsistent und stabil sein.

In mehreren Untersuchungen ist aufgezeigt worden, daß Antwortunterschiede in Abhängigkeit von der Fragenformulierung vor allem auftreten, wenn die Vpn von einem Sachverhalt nicht betroffen, an ihm nicht interessiert oder über ihn nicht informiert sind (vgl. vor allem Payne 1951, Noelle-Neumann 1970, Rugg & Cantril 1972). Sicherlich ist es in solchen Fällen unsinnig, aus Antworten, die weitestgehend methodenbedingt sind, einen Inhalt herausmitteln zu wollen. Angemessener wäre es wohl zu folgern, daß es fast ausschließlich von der Formulierung der Frage abhängt, was die Vpn antworten.

Ausgehend von den Überlegungen von Campbell & Fiske (1959) sollte man gewissermaßen im Rahmen eines ‚Multi-Content-Multi-Question-Ansatzes‘ für Fragen konvergente und diskriminante Validität fordern und Antworten erst dann interpretieren, wenn diese Forderungen erfüllt sind. Konkret würde das bedeuten, daß unterschiedlich formulierte Fragen zu einem bestimmten Inhalt zu ähnlichen Antworten führen müssen (Methodenkonvergenz), mindestens aber zu ähnlicheren Antworten als vergleichbar formulierte Fragen zu verschiedenen Inhalten (diskriminante Validität). Hält man sich vor Augen, daß allein durch Formulierungsähnlichkeit etwa auf dem Wege über die Wirkung von response sets hohe Korrelationen zwischen Antworten auf verschiedene Fragen zustandekommen und sich z.B. in einer Faktorenanalyse als ‚Formulierungsfaktor‘ niederschlagen können (Holm 1974a, b), wird man auch diskriminante Validität nicht mehr einfach unterstellen können, wie das heute noch vielfach geschieht.

4. Reihenfolge der Fragen und Umfang des Fragebogens

In stärkerem Maße noch als das bei der Formulierung von Fragen der Fall ist, stützt sich der Fragebogaufbau üblicherweise auf Vermutungen und unsystematische Erfahrungen von Praktikern (Bradburn & Mason 1964). Die relativ wenigen empirischen Untersuchungen über Auswirkungen der Fragenreihenfolge und des Fragebogensumfangs können nur begrenzte Gültigkeit beanspruchen, so daß die Feststellung von Kreutz & Titscher (1974, 40), derzufolge „...über den Aufbau des Fragebogens sehr wenig gesichertes Wissen vorhanden ist“, auch heute noch zutreffen dürfte.

4.1 Ziele beim Aufbau eines Fragebogens

Aus prinzipiell den gleichen Gründen, wie sie für die Festlegung von Fragenformulierungen angeführt wurden (vgl. 3.1.1), scheint in vielen Fällen auch die Standardisierung der Fragenfolge der mit dem geringeren Risiko für Verzerrungen behaftete Weg zu sein. Allerdings schließt dieser Weg naturgemäß die im freien mündlichen Interview gegebene Möglichkeit der Anpassung der Fragenfolge an die Erfordernisse der jeweiligen Befragungssituation durch den Interviewer aus. Damit sind prinzipiell Gefahren für die Motivation des Befragten verbunden, z.B. wenn ihm eine an früherer Stelle unaufgefordert bereits beantwortete Frage entsprechend ihrer Position im Fragebogen später erneut gestellt wird (Noelle 1963). Um diese und ähnliche Schwierigkeiten möglichst zu vermeiden, formulieren Karmasin & Karmasin (1977, 197) als Leitlinie für den Aufbau eines Fragebogens, diesen „... so zu gestalten, daß für den Befragten der Charakter eines Gesprächs, einer Konversation simuliert wird“. Ähnlich äußern sich auch Kirschhofer-Bozenhardt & Kaplitza (1975). Praktisch bedeutet dies u.a., daß Fragen, die zusammenhängen, auch im Zu-

sammenhang zu stellen sind, zumal auf seiten der Vpn ein ausgeprägtes Bedürfnis zu bestehen scheint, Zusammenhänge zwischen Fragen bzw. Frageninhalten herzustellen (Karmasin & Karmasin 1977). Die Forderung nach Gruppierung des Zusammengehörigen kollidiert möglicherweise mit der den gebräuchlichen Meßmodellen zugrundeliegenden Annahme stochastischer Unabhängigkeit der Antworten auf verschiedene Fragen (Wottawa 1980), d.h. alle Arten von Reihenfolgeeffekten (seien sie unbeabsichtigt oder, wie bei Fragen, die auf einen bestimmten Sachverhalt hinführen sollen, bewußt eingesetzt), stellen einen Verstoß gegen Grundannahmen der Meßmodelle (Unkorreliertheit von Zufallsfehlern bzw. Abhängigkeit der Antworten nur von Item- und Personenparametern) dar. Aus diesem Grund ist - neben der ‚Natürlichkeit‘ des Gesprächsverlaufs - die Ausschaltung von Reihenfolgeeffekten (d.h. von Einflüssen auf die Antworten, die sich allein aus der Reihenfolge der Fragen ergeben) ein mit dem erstgenannten nur mehr oder weniger zu vereinbarendes Ziel des Fragebogaufbaus.

Noelle-Neumann (1974) führt daneben die Motivierung der Befragten (d.h. das Bemühen, Interesse für die Befragung bzw. die Frageninhalte zu wecken) und die ‚Optimierung der Auskunftsfähigkeit‘ (d.h. die Steigerung bzw. Aufrechterhaltung der Aufmerksamkeit über den Befragungsverlauf) als wichtige Ziele an, die bei der Festlegung der Reihenfolge von Fragen im Auge zu behalten sind. Die Orientierung am natürlichen Verlauf eines Gesprächs kann in dieser Richtung wirken, ist mit diesen Zielen jedoch nicht identisch. Im Interesse der Verminderung interindividueller Beantwortungsunterschiede (Fehlervarianz im Falle der demoskopischen Befragung) und der Interpretierbarkeit von Subgruppenergebnissen müssen außerdem eine ‚Vergleichbarkeit des Befragungsablaufs‘ (z.B. im Falle von Verzweigungen) für alle Vpn angestrebt und bei der Festlegung der Fragenfolge auch die spätere Auswertbarkeit (Belange der Datenerfassung) im Auge behalten werden (Noelle-Neumann 1974).

Kontrovers behandelt wird die Frage, ob dem ‚natürlichen‘ bzw. ‚logischen‘ Aufbau des Fragebogens (der Fragenfolge) ein Wert an sich beizumessen (bzw. er infolge Strukturierungsbedürfnisses auf seiten der Vpn unvermeidbar) sei (z.B. Phillips 1966, Cannell et al. 1977, Karmasin & Karmasin 1977) oder ob ein ‚logischer‘ Fragebogaufbau allenfalls ein denkbare Mittel unter vielen auf dem Weg zur Motivierung von Vpn, Verbesserung ihrer Auskunftsfähigkeit, Vermeidung von Reihenfolgeeffekten und Sicherstellung der Vergleichbarkeit des Befragungsablaufes darstelle (Stroschein 1965, Noelle-Neumann 1974). Entsprechend unterscheiden sich die Autoren auch darin, welchen Stellenwert sie dem ‚Themenwechsel‘ im Aufbau eines Fragebogens einräumen.

Bei der Abfolge von Fragen muß grundsätzlich unterschieden werden zwischen der Abfolge von Frageninhalten (‚Themendisposition‘ im Sinne von Stroschein 1965) und von Fragentypen (‚Fragendisposition‘, vgl. Stroschein 1965).

4.2 Motivation der Befragten und Steigerung der Antwortfähigkeit

Einerseits können sich durch Veränderungen der Motivation der Befragten im Verlaufe des Interviews Einflüsse auf Antworten in Abhängigkeit von der Position der Frage ergeben (vgl. 4.3.2, 4.4 und 4.5), insofern stellt die Motivation der Befragten eine mögliche Ursache beobachtbarer Reihenfolgeeffekte dar (Anger 1969). Andererseits ist die Gestaltung der Fragenfolge aber ein Mittel, angemessene Motivation der Befragten zu erreichen bzw. zu erhalten (vgl. besonders Perreault 1975). Hierzu schlägt Noelle-Neumann (1974, vgl. auch Noelle 1963) vor, besonderes Augenmerk den einleitenden Fragen zu schenken und diese (notfalls als ‚Wegwerf-Fragen‘) zu Kontakt- bzw. ‚Eisbrecher‘-Fragen zu machen, die insbesondere mißtrauischen und unsicheren Vpn (z.B. älteren Menschen, Angehörigen der Unterschicht, Hausfrauen) ‚Sicherheit‘ vermitteln sollen. Als geeignete Themen gelten z.B. die erwartete Entwicklung der Preise, der Einfluß des Wetters auf die Befindlichkeit u.ä., wobei die Fragen zwar leicht zu beantworten und nicht kontrovers (Goode & Hatt 1972), andererseits aber auch nicht banal sein sollten. Um die Vpn „ins Gespräch zu ziehen“ (Noelle-Neumann 1974, 244), empfiehlt sich evtl. eine offene Frage.

Während des Interviews sollen sowohl Motivation (Antwortbereitschaft) als auch Antwortfähigkeit (die sich vermutlich nicht streng trennen lassen) durch Wechsel der Themen, Wechsel der Inhalte (Wissen, Fakten, Meinungen, Verhalten) und Wechsel der Fragentypen (geschlossene, offene Fragen, Listen- oder Kartenvorlagen, wechselnde Formate und Farben des Vorlagematerials) aufrechterhalten bzw. gesteigert werden (Stroschein 1965). Noelle-Neumann (1963, 1974) schlägt sogar einen eigenen Typ instrumenteller Fragen, die sogenannten ‚Spielfragen‘ (Beurteilung von Frisuren, Kleidern, Farbwahlen etc.) nur zur Beeinflussung von Motivation bzw. Aufmerksamkeit vor. Lange Serien geschlossener Fragen gelten als frustrierend und monotoniefördernd, Serien offener Fragen als anstrengend und dadurch ermüdend (Noelle-Neumann 1974).

Die behaupteten Wirkungen spezieller Einleitungsfragen und der verschiedenen Techniken zur Beeinflussung von Aufmerksamkeit und Motivation sind empirisch allerdings nicht abgesichert (Kreutz & Titscher 1974).

Im Zusammenhang mit der Forderung nach häufigem, durchaus auch sprunghaftem (Noelle 1963) Themenwechsel werden von Autoren aus dem Bereich der kommerziellen Markt- und Meinungsforschung die sogenannten Mehrthemenumfragen (Omnibus-Befragungen) als methodisch besonders vorteilhaft hervorgehoben. Es stellt sich allerdings die Frage, ob hier nicht eine Not zur Tugend gemacht werden soll. Immerhin betont Richter (1969), daß jeder The-

menwechsel mit einer besonderen Anstrengung für den Befragten (Umorientierung) verbunden sei, und fordert für unpersönlich-schriftliche (postalische) Befragungen im Interesse eines hohen Rücklaufs eine Zusammenstellung von Fragen nach Maßgabe der Befragungsthemen zu sogenannten ‚assoziativen Blöcken‘. Eine möglichst sinnvolle Ordnung von Fragen wird von Autoren wie Phillips (1966), Goode & Hatt (1972), (mit Einschränkungen) Holm (1974b), Cannell et al. (1977), Karmasin & Karmasin (1977) gefordert (vgl. auch 4.1). Auch Anger (1969) warnt vor zu starkem Themenwechsel, von dem er Gefahren für die erlebte Seriosität der Befragung ausgehen sieht. Empirische Untersuchungen, in denen Themenwechsel und logischer Fragebogaufbau verglichen worden wären, scheinen nicht zu existieren (vgl. Kreutz & Titscher 1974), allerdings ist auch zweifelhaft, ob sie generalisierbare Befunde zutage fördern könnten. Vermutlich gibt es nur den Weg, unter Berücksichtigung von möglichen Monotonieeffekten einerseits und Belastungen durch inhaltliche Umorientierung sowie des Bedürfnisses der Vpn nach sinnvollem Zusammenhang der Fragen andererseits, die günstigste Themen- und Fragendisposition im Einzelfall durch Pretest empirisch zu bestimmen.

4.3 Reihenfolgeeffekte

Einflüsse der Stellung einer Frage innerhalb eines Fragebogens auf die Antworten können einmal durch die Inhalte vorangegangener Fragen, dann aber auch unabhängig von diesen Inhalten dadurch zustande kommen, daß die Frage früher oder später im Verlauf einer Befragung gestellt wird und das Antwortverhalten der Vpn sich über die Dauer der Befragung verändert (Bradburn & Mason 1964). Da vorausgehende Fragen immer Inhalte haben und andererseits das Vorausgehen einer Frage bestimmten Inhaltes die zu betrachtende Frage notwendig an eine spätere Stelle verschiebt, ist es prinzipiell nicht möglich, diese Effekte völlig voneinander zu isolieren. Im Interesse theoretischer Klarheit werden dennoch im folgenden als ‚Kontexteffekte‘ Einflüsse des Inhaltes vorangehender Fragen und als ‚Positionseffekte‘ Einflüsse der relativen Position auf die Antworten zu einer gegebenen Frage unterschieden.

4.3.1 Kontexteffekte

Diese oft auch als Ausstrahlungseffekte angesprochenen Einflüsse vorangegangener auf nachfolgende Fragen lassen sich nach einem auf Bradburn & Mason (1964) zurückgehenden Vorschlag in

- Aktualisierungs- (Präsenz-, saliency-) Effekte oder allgemeiner Lerneffekte (Anger 1969),
 - Konsistenzeffekte und
 - Redundanzeffekte
- einteilen.

Aktualisierungseffekte kommen dadurch zustande, daß eine vorausgegangene Frage die Antworten auf eine nachfolgende beeinflusst, indem sie bestimmten Sachverhalten oder bestimmten Bezugsrahmen im Bewußtsein der Befragten höheres Gewicht verleiht. Hierfür finden sich in der Literatur mehrfach Beispiele bzw. empirische Belege. Stellt man z.B. zunächst eine Frage nach Erwartungen zur Preisentwicklung und danach eine solche nach den wichtigsten Fragen, mit denen Politiker sich in nächster Zeit beschäftigen sollten, ist zu erwarten, daß die Preisstabilität erheblich häufiger genannt wird, als sie ohne eine derartige vorausgegangene Frage genannt worden wäre (Noelle-Neumann 1974), einfach weil die Preisstabilität als politisches Thema einen höheren Grad der Bewußtheit erhalten hat. Durch Aktualisierung eines geeigneten Bezugsrahmens läßt sich erklären, daß ‚Kartoffeln‘ in einer Untersuchung, die Noelle-Neumann (1970) referiert, von 30% der Befragten die Eigenschaft eines ‚deutschen‘ Nahrungsmittels zugesprochen wurde, wenn nach ihnen vor, von 48%, wenn nach ihnen h i n t e r dem Nahrungsmittel ‚Reis‘ gefragt wurde (ähnliche Reihenfolgeeffekte gab es auch für Reis und Nudeln). Willick & Ashley (1971) befragten College-Studenten, welche politischen Parteien sie und welche ihre Eltern bevorzugen würden, und erhielten signifikant mehr übereinstimmende Angaben (für die eigene Bevorzugung und die der Eltern), wenn zuerst nach der Bevorzugung des Studenten und danach nach der der Eltern gefragt wurde. Sie erklären diesen Befund mit dem Bemühen der Studenten, Unabhängigkeit von den Ansichten ihrer Eltern zu demonstrieren. Dies war dann nicht ohne weiteres möglich, wenn die Studenten zum Zeitpunkt der Antwort betreffend ihre eigene Meinung nicht wußten, daß sie auch nach der Haltung ihrer Eltern (die für sie in der Regel anschaulich festlag) befragt werden würden. Weitere Beispiele beziehen sich auf das Recht des Eintritts für Amerikaner in die deutsche bzw. englische oder französische Armee während des Zweiten Weltkrieges (Rugg & Cantril 1972) und auf die Haltung von Einwohnern der BRD gegenüber den USA bzw. der UdSSR (Noelle-Neumann 1970).

Aktualisierung kann je nach Befragungszielen ein unerwünschter, evtl. aber auch ein erwünschter Reihenfolgeeffekt sein. Besteht das Ziel der Befragung darin, Beurteilungen oder Bewertungen von Sachverhalten möglichst unbeeinflusst von Aktualisierungen zu erhalten, wählt man häufig eine als ‚Trichter‘ bezeichnete, vom Allgemeineren zum Spezielleren fortschreitende Reihenfolge (vgl. Maccoby & Maccoby 1972, Hennig 1975, Karmasin & Karmasin 1977). Friedrichs (1973) beschreibt die von Gallup verwendete Standard-Fragenfolge eines Trichters:

- Vertrautheit mit dem Sachverhalt
(offene Wissensfrage, z.B. ‚Was verstehen Sie unter . . .‘),
- unbeeinflusste Einstellung
(offene Einstellungsfrage, z.B. ‚Was sollte X für . . . tun?‘),

- Reaktion auf spezifische vorgegebene Einstellungen
(geschlossene Fragen, z.B. „Manche sagen . . . andere sagen . . . was meinen Sie ist richtig?“),
- Begründung der Reaktion auf vorgegebene Einstellungen
(offene Warumfrage),
- Intensität der Einstellung
(Skalafrage).

In Fällen, in denen die aktualisierende Wirkung vorausgegangener Fragen befragungstaktisch erwünscht bzw. erforderlich ist, bedient man sich gelegentlich auch der Technik des umgekehrten Trichterns, d.h. des Fortschreitens vom Speziellen, Konkreten zum Allgemeinen, Abstrakten (Maccoby & Maccoby 1972). Hennig (1975) führt z.B. aus, daß eine Frage an Arbeiter nach der vorausgesehenen Organisation von Produktionsabläufen in zehn Jahren kaum zu verwertbaren Antworten führen dürfte, wenn sie ‚unvermittelt‘ gestellt wird. Erfolgsversprechender ist hier eine Fragenfolge nach Art eines umgekehrten Trichters, z.B.

- ‚Sind in Ihrem Betrieb in der nächsten Zeit Neuerungen im Produktionsablauf geplant? Wenn ja: welche?‘,
- ‚Was glauben Sie, wie in zehn Jahren der Produktionsablauf aussehen wird?‘.

Befragungstaktisch beabsichtigt und gezielt eingesetzt werden Aktualisierungen durch geeignete Fragenfolge auch, wenn seitens der Vp eine Reproduktion von Gedächtnisinhalten erforderlich ist. Cannell & Kahn (1968) schlagen in solchen Fällen vor, die Vpn z.B. durch chronologisch geordnete Fragen auf den thematischen Sachverhalt hinzuführen (vgl. auch Mauldin & Marks 1950, Phillips 1966).

Von *Konsistenzeffekten* der Fragenfolge spricht man, wenn die Vp eine Frage nicht ‚zutreffend‘ sondern so beantwortet, daß sie zu ihren Antworten auf vorangegangene Fragen nicht in Widerspruch gerät. Noelle-Neumann (1974) nennt als Beispiel Aussagen von Befragten über Aufwendungen für ‚Luxusartikel‘ (z.B. Blumen), die dann niedriger angegeben werden, wenn die Befragten in einer vorangegangenen Frage ‚sparsame Lebensführung‘ für sich in Anspruch genommen haben. Auch Holm (1974 b) weist auf solche Gefahren hin, insbesondere dann, wenn die Fragen zu einer bestimmten ‚Zieldimension‘ gruppiert (zusammengestellt) sind. Bradburn & Mason (1964) gelang es andererseits nicht, in ihren Untersuchungen Anhaltspunkte für derartige Reihenfolgeeffekte zu finden: Antworten auf eine Frage nach ‚globaler‘ Zufriedenheit wurden hier nicht davon beeinflusst, ob Fragen zur Zufriedenheit mit speziellen Aspekten vorausgegangen waren oder nicht.

Ist mit ausgeprägten Konsistenzneigungen der Vpn zu rechnen, so empfiehlt sich nach Noelle-Neumann (1974) eine Fragenfolge nach Art des umgekehrten

Trichters: Spezifische Angaben (z.B. ‚Ausgaben für Luxusartikel‘) würden dann als weniger widersprüchlich mit allgemeinen („prinzipielle Sparsamkeit“) empfunden, wenn sie diesen vorangehen. Mit ähnlicher Begründung empfehlen Tittle & Hill (1967) erst nach *Verhalten* und dann nach *Einstellungen zu fragen*.

Als *Redundanzeffekt* bezeichnen Bradburn & Mason (1964) das Ausbleiben bestimmter Antworten auf Fragen dadurch, daß diese Antworten bereits auf frühere Fragen gegeben wurden und die Vpn sich nicht wiederholen wollen. Sie berichten von geringeren Häufigkeiten für die Nennung bestimmter (z.B. Partner-)Probleme in einer offenen Frage nach ‚Sorgen‘, wenn diese Probleme bereits Gegenstand vorangegangener Fragen waren. Auch Noelle-Neumann (1974) betont, daß man in solchen Fällen die Antworten auf die spätere Frage nicht unabhängig von denen auf die vorangegangenen Fragen betrachten und behandeln dürfe.

In gewissem Sinne liegen Reihenfolgeeffekte nach Art von Kontexteffekten auch vor, wenn durch (erfolgreiche) Verwendung von Puffer- oder Ablenkungsfragen (vgl. 2.1) Auswirkungen früherer auf spätere Fragen *vermieden* werden: Auch hier lauteten die Antworten anders, würden diese instrumentellen Fragen der thematischen Frage nicht vorangehen. Darüber hinaus lassen sich durch die Schwierigkeitsabstufung von Fragen Reihenfolgeeffekte erzeugen, etwa im Sinne einer Erleichterung besonders schwieriger Fragen durch langsamen Schwierigkeitsanstieg oder im Sinne einer Überwindung von ‚Antworthemmungen‘ durch starke Schwierigkeitsunterschiede (vgl. 4.4).

Daß im Falle unpersönlicher schriftlicher (postalischer) Befragungen infolge Nichtkontrollierbarkeit der Reihenfolge der Bearbeitung von Fragen auch Auswirkungen nachfolgender auf vorangehende Fragen möglich sind, sei der Vollständigkeit halber erwähnt. Entsprechendes gilt, da der Interviewer ja den ganzen Fragebogen kennt und die registrierten Antworten auf verschiedenen Wegen mitbeeinflußt, übrigens auch für mündliche Interviews.

4.3.2 Positionseffekte

Nach Richter (1969) und Goode & Hart (1972), die dafür allerdings empirische Belege nicht vorlegen, nimmt die Wahrscheinlichkeit für den Abbruch eines Interviews vom Anfang zum Ende des Interviews hin ab. Karmasin & Karmasin (1977) führen dies (im Falle von mündlichen Interviews) darauf zurück, daß mit der Interaktionshäufigkeit auch die Sympathie zwischen Interviewer und Befragtem (man muß wohl ergänzen: in der Regel) wachse, und leiten daraus auch eine Tendenz zu *weniger negativen* Urteilen an späterer Stelle im Interview ab. Kraut et al. (1975) konnten empirisch allerdings nur leichte Tendenzen in Richtung auf *weniger extreme* Urteile und mehr Auslassungen

bzw. Nichtbeantwortungen gegen Ende einer persönlichen schriftlichen Befragung nachweisen. Von abnehmender ‚Sorgfalt‘ im Verlauf des Interviews berichtet auch Stroschein (1965). Johnson et al. (1974) kamen in entsprechenden Untersuchungen für eine offene Frage zu dem Ergebnis, daß diese insgesamt am meisten Information lieferte, wenn sie einmal am Anfang und dann erneut am Ende einer Serie von 18 bzw. 62 geschlossenen Fragen gestellt wurde. Besteht (wie im Regelfall) nur die Möglichkeit, die Frage einmal zu stellen, so liefert sie am Anfang des Fragebogens mehr Information als am Ende, d.h. der mögliche Zugewinn an Aspekten durch Lernprozesse während der Befragung wird durch Effekte verringerter Motivation bzw. Aufmerksamkeit überkompensiert.

Zusammenfassend können die vorliegenden Befunde wohl als Hinweise darauf gelten, daß insbesondere schwierige und anstrengende Fragen nicht zu spät, wegen der Abbruchgefahr aber auch nicht zu früh im Fragebogen auftauchen sollten. Beurteilungen sind in ihren extremen Ausprägungen nur vergleichbar, wenn die Fragen etwa gleiche Positionen im Fragebogen hatten, betrachtet man allerdings nicht die Extremkategorien (z.B. ‚sehr dafür‘), sondern die Mittelwerte der Einstufungen, so sind (mindestens nach den Befunden von Kraut et al. 1975) Positionseffekte kaum noch zu befürchten.

4.4 Unangenehme und heikle Fragen

übereinstimmend findet sich in der Literatur die Empfehlung, unangenehme bzw. heikle Fragen (z.B. solche nach Einkommen, Kindererziehung, Allgemeinbildung, Sexualität, Familienverhältnissen, körperlicher Sauberkeit, vgl. 2.3) erst in der zweiten Hälfte des Fragebogens zu stellen, um einerseits das verringerte Risiko für Abbrüche, andererseits das angewachsene ‚Vertrauen‘ der Befragten zu nutzen (Kreutz & Titscher 1974, Karmasin & Karmasin 1977). Darüber hinaus gibt es jedoch spezielle Techniken der Berücksichtigung heikler Fragen im Fragebogaufbau. Im Sinne einer Erfahrungsregel schlägt Noelle-Neumann (1974) z.B. vor, solche Fragen besonders einfach zu formulieren und sie nach betont schwierigen Fragen (z.B. offenen Wissensfragen mit schwierigen Inhalten) zu platzieren. Durch einen Kontrasteffekt werde die kritische Frage dann als besonders leicht erlebt und gewissermaßen nach Art einer Selbstüberrumpelung beantwortet, bevor auf Seiten der Vp mögliche Antworthemmungen überhaupt zur Wirkung kommen könnten. Koolwijk (1968) fand in seinen Untersuchungen Angleichungs- und Kontrasteffekte für die Unangenehmheit von Fragen: Ließ er eine unangenehme Frage auf eine neutrale folgen, so verstärkte sich die Unangenehmheit (Kontrast), ging umgekehrt die unangenehme Frage der neutralen voran, so wurde tendentiell auch letztere als unangenehm erlebt (Angleichung, halo). Aus diesen Befunden leitet er die Forderung ab, einer inhaltlich interessierenden unangenehmen Frage

zur ‚Einstimmung‘ und zur Vermeidung des Kontrasteffektes eine ebenfalls unangenehme (evtl. Wegwerf-Frage) voranzuschicken und nachfolgende neutrale Fragen durch Pufferfragen gegen vorausgegangene unangenehme Fragen abzuschirmen.

Schließlich haben Goode & Hatt (1972) darauf hingewiesen, daß Unangenehmheit nicht nur eine Einzelfrage, sondern auch eine Fragenfolge kennzeichnen kann, d.h. sie fordern für den Fragebogaufbau die Vermeidung von Fragenfolgen, die für bestimmte Vpn peinlich werden könnten. Der Fragenfolge 1. ‚Haben Sie Kinder?‘, 2. ‚Sind Sie verheiratet?‘ wäre unter diesem Kriterium die umgekehrte Folge mit Filterung ‚Sind Sie verheiratet?, wenn ja: Haben Sie Kinder?‘ vorzuziehen.

4.5 Fragen zur Person

Die unsicheren Grundlagen einer an Erfahrungsregeln statt an empirischen Untersuchungsergebnissen orientierten Fragebogenkonstruktion werden im Zusammenhang mit den Empfehlungen deutlich, die verschiedene Autoren für die Position von Angaben zur Person (biographischen Angaben bzw. demographischen Fragen) geben. So stellt Noelle-Neumann (1974, 244) nachdrücklich fest: „Personenstandsdaten gehören nicht an den Anfang des Interviews, sondern an das Ende; an den Anfang gesetzt geben Sie dem Interview den Charakter eines Verhörs“. Eine ähnliche Position vertritt auch Stollberger (1966). Dem steht die Auffassung von Kreutz & Titscher (1974) entgegen, die Fragen zur Person an den Anfang des Fragebogens gestellt wissen möchten, weil ihnen einerseits die Forderung nach der Schlußposition für solche Fragen empirisch nicht begründet zu sein scheint und sie andererseits sorgfältigere und damit gültigere Antworten erwarten, wenn die Vp gleich zu Beginn der Befragung (bei den Fragen zur Person) feststellt, daß sich der Untersucher für sie als Individuum interessiert. Auch hier handelt es sich allerdings um eine Spekulation, die man im Vergleich zur Gegenposition für plausibler halten kann oder auch nicht.

4.6 Filterfragen und Verzweigungsfragen

Ablauf-Ordnungsfragen (vgl. 2.1) wie Filter- und Gabelungs- bzw. Verzweigungsfragen stellen im Interesse der Vermeidung von ‚Unterstellungen‘ (vgl. 3.1.4) bzw. der Nichtbelastung von Vpn mit unzutreffenden Fragen ein häufig unverzichtbares Mittel bei der Festlegung einer angemessenen Fragenfolge innerhalb eines Fragebogens dar. Andererseits sollte sich der Fragebogenkonstrukteur vergegenwärtigen, daß mit dem Einbau von Filterungen und Verzweigungen in einen Fragebogen die Fehlerhäufigkeit unweigerlich ansteigt.

So berichten Cannell et al. (1977) für mündliche standardisierte Interviews, daß

- von Fragen, die allen Vpn gestellt werden sollten, nur 1,5% bei mehr als 10% der Vpn ausgelassen wurden, aber
- von Fragen, die infolge Filterung bzw. Verzweigung nur für Subgruppen vorgesehen waren, 54% bei mehr als 10% der Vpn nicht oder nicht richtig gestellt wurden.

Richter (1969) hat für den Fall unpersönlich-schriftlicher (postalischer) Befragung die Beachtung von Filteranweisungen untersucht und je nach Bildungsstand, Einkommen, Beruf etc. Nichtbeachtungsanteile bis in die Größenordnung von 30% der Befragten gefunden, so daß die Verwendung von Filterungen und Gabelungen beim Aufbau von Fragebogen für unpersönlich-schriftliche Befragung möglichst zu vermeiden ist (Wieken 1974).

4.7 Spezielle Gesichtspunkte für die Itemreihenfolge diagnostischer Fragebogen

Für diagnostische Fragebogen stellt sich insbesondere die Frage, ob Items nach ihrer ‚Zieldimension‘ gruppiert bzw. ob sie in Zufallsfolge vorgegeben werden sollten. Die Gruppierung von Items nach ihrem Inhalt erhöht tendenziell die Durchschaubarkeit und wird daher für empirische Konstruktionen (vgl. 1.1.2) von vornherein nicht in Betracht gezogen. Hier liegen prinzipiell zufällige Itemfolgen vor, die allenfalls im Interesse leichterer Auswertbarkeit durch für die Vpn nicht ersichtliche systematische Anordnungen durchbrochen werden (vgl. als Beispiel den MMPI, die ‚Lügenitems‘ sind hier systematisch angeordnet, Hathaway & Mc Kinley 1963).

Sieht man von der höheren Durchschaubarkeit und damit Verfälschbarkeit und der Störung der stochastischen Unabhängigkeit der Einzelantworten ab, so kann für Fragebogen mit inhaltlichem Validitätsanspruch eine inhaltliche Gruppierung von Items zu einer ‚Sensibilisierung‘ der Vp in dem Sinne führen, daß sie durch die Zusammenstellung der Items ihre ‚Lage‘ auf der Zieldimension besser bestimmen kann, als sie das anhand verstreuter Items tun könnte. Dies würde zu valideren Antworten führen. Andererseits kann die Zusammenstellung von Items die Vp aber auch zu inadäquat konsistentem Antwortverhalten veranlassen und dadurch die Validität beeinträchtigen. Es ist nur im Einzelfall zu klären, ob vorwiegend die (erwünschte) Sensibilisierungstendenz‘ oder die (unerwünschte) ‚Konsistenztendenz‘ (Holm 1974 b) durch eine inhaltliche Gruppierung der Items begünstigt wird. Daß eine solche Gruppierung nicht notwendig zu artifizieller Konsistenz im Antwortverhalten führen muß, haben Metzner & Mann (1953) gezeigt: Sie fanden keine systematische Erhöhung der Interkorrelationen von Items durch Gruppierung; für Einzelfäl-

le berichten sie sogar erhebliche Senkungen der Korrelationen zwischen Items. Sie begründen dies einleuchtend mit itemspezifischen Kontexteffekten, d.h. die Bedeutung eines Items kann sich durch direkte Nachbarschaft zu anderen Items ändern und zwar nicht nur (wie unter der Konsistenz-Hypothese erwartet) in Richtung auf höhere (Un-)Ähnlichkeit zu den anderen Items der entsprechenden Dimension.

Darüber hinaus sind bei diagnostischen Fragebogen die Effekte der Reihenfolge der Items abhängig von der ‚Sicherheit‘ bzw. ‚Zugänglichkeit‘ des zu erfassenden Sachverhaltes bzw. der Lage einer Vp auf dem interessierenden latenten Kontinuum. Hayes (1964) konstruierte je eine Guttman-Skala aus ‚Angst-Items‘ und ‚Mathematik-Aufgaben‘ und stellte fest, daß zwar bei Angst-Items, nicht aber bei Mathematik-Aufgaben die (in aufsteigender Schwierigkeit) geordnete Vorgabe der Items zu signifikant anderen Antworten und damit (geringeren) Angst-Werten führte als die ungeordnete Vorgabe. Eingeschobene ‚irrelevante‘ Items blieben bei beiden Skalen ohne Auswirkungen.

4.8 Überlegungen zur Vermeidung unerwünschter Reihenfolgeeffekte

Für die Elimination von Reihenfolgeeffekten aus den Ergebnissen für Vpn-Gruppen gelten prinzipiell diejenigen Überlegungen, die im Zusammenhang mit der Kontrolle von Einflüssen der Fragenformulierung in 3.4 angestellt worden sind. Wie dort, so kann auch bezüglich der Fragenreihenfolge im Rahmen einer gegabelten Befragung (Split-ballot-verfahren) mit verschiedenen Fragebogenvarianten gearbeitet und die Absicht verfolgt werden, die Reihenfolgeeffekte einflüsse ‚herauszumitteln‘ (vgl. zu dieser Vorgehensweise die Ausführungen über die Variation der Reihenfolge von Antwortvorgaben in 3.1.3). Wie im Falle der Formulierungseffekte muß aber auch hier gefragt werden, ob solche ‚mittlere‘ Antworten eine inhaltliche Bedeutung haben oder ob die Existenz von Reihenfolgeeffekten ein Hinweis darauf ist, daß die Erfassung des Inhaltes mit der verwendeten Methode nicht oder nur unzureichend gelingt.

Hält man die Variation der Fragenreihenfolge für ein angemessenes Verfahren, so ist es prinzipiell günstig, mit möglichst vielen Varianten des Fragebogens zu arbeiten und im Extremfall für jeden Befragten einen eigenen Fragebogen zu erstellen. Die damit verbundenen Probleme und die Möglichkeiten, die der Einsatz elektronischer Datenverarbeitungsanlagen zur Erstellung auswertbarer individualisierter Fragebogen bietet, diskutiert Perreault (1976; für den Fall des semantischen Differentials vgl. auch Kane 1969). Cataldo et al. (1970) schlagen nicht zuletzt wegen leichter Variierbarkeit der Reihenfolge den verstärkten Einsatz von card-sorting-Techniken im Rahmen von (persönlichen) Befra-

ungen vor: Statt für die einzelnen Statements Fragen nach dem Grad ihres Zutreffens beantworten zu lassen, werden dabei die Statements auf Karten geschrieben und den Vpn zur Einordnung in bestimmte Antwortkategorien (z.B. ‚sehr dafür‘, ‚dafür‘, ‚dagegen‘, ‚sehr dagegen‘) übergeben.

Soweit man die Variation der Reihenfolge nicht für ein angemessenes Vorgehen hält, mit der Existenz von Reihenfolgeeffekten aber rechnen muß, ist entsprechend den Ergebnissen von Kraut et al. (1975; vgl. auch 4.3.2) zu berücksichtigen, daß streng vergleichbar nur Antworten auf Fragen sind, die in Fragebogen an der gleichen Position (und im gleichen Kontext) verwendet wurden, und daß die Antworten außer inhaltlichen auch Reihenfolgeeffekte widerspiegeln. Je intensiver, klarer, sicherer die zu erfassenden Sachverhalte für die Vpn sind, desto geringer sind ceteris paribus die Einflüsse der Fragenfolge auf die Antworten (vgl. Bradburn & Mason 1964, Hayes 1965, Willick & Ashley 1971).

4.9 Fragebogenumfang

Zwar findet sich in fast jeder Darstellung von Befragungsmethoden auch eine Angabe bzw. Empfehlung bezüglich des akzeptablen Fragebogenumfanges bzw. der akzeptablen Interview- bzw. Bearbeitungsdauer, doch handelt es sich dabei stets nur um Erfahrungswerte bzw. common-sense-Angaben. Typisch dafür ist z.B. Noelle (1963), die als Richtwert für die Dauer eines mündlichen Interviews 30 Minuten nennt (ähnliche Werte finden sich z.B. auch bei Kirschhofer-Bozenhardt & Kaplitza 1975, Karmasin & Karmasin 1977), bei ‚gutem Aufbau‘ aber auch mehr als eine Stunde für möglich hält. Als Indikator für die Einhaltung bzw. Überschreitung der akzeptablen Dauer schlägt sie vor, die Befragten am Ende des Interviews diese Dauer schätzen zu lassen: Werde sie unterschätzt, sei das Interview nicht zu lang gewesen.

Empirische Untersuchungen der Wirkung unterschiedlicher Fragebogenumfänge wurden - soweit bekannt - nur im Zusammenhang mit unpersönlich-schriftlichen (postalischen) Befragungen durchgeführt, wobei abhängige Variable stets allein die Rücklaufquote war. Berdie (1973), der auch ältere Untersuchungen referiert und die prinzipiell plausible negative Korrelation zwischen Fragebogenumfang und Rücklaufquote, von der viele Autoren berichten, als tradiertes ‚Einvernehmen‘ ohne nennenswerte empirische Basis entlarvt, fand zwar unterschiedliche Rücklaufquoten von 64%, 56% und 42% für Fragebogen mit einer Seite (10 Fragen), 2 Seiten (20 Fragen) und 4 Seiten (40 Fragen), doch waren diese Unterschiede (bei 108 Vpn) statistisch nicht bedeutsam. Sheth & Roscoe (1975) verglichen die Rücklaufquoten für einen vierseitigen (23 Items, 10 min Bearbeitungszeit) und einen sechsseitigen (49 Items, 18 min Bearbeitungszeit) Fragebogen und fanden keinen Unterschied, wobei aller-

dings zu bedenken ist, daß die Fragebogenumfänge nur wenig differierten und der längere Fragebogen sich vom kürzeren auch inhaltlich systematisch unterschied. Unterhalb des von ihm untersuchten Maximums von 4 Seiten und einer Bearbeitungszeit von 15-20 min fand auch Richter (1969) keinen Zusammenhang zwischen Fragebogenumfang und Rücklaufquote, und Perreault (1975) berichtet von sehr hohen Rücklaufquoten sogar für einen 9 Seiten umfassenden Fragebogen, der allerdings ‚personalisiert‘ (mit persönlich wirkendem Anschreiben etc. versehen) war (vgl. auch Erdos 1970, Linsky 1975).

Natürlich ist mit diesen Untersuchungen nicht bewiesen, daß es keinen Einfluß des Fragebogenumfangs auf Rücklaufquoten gibt, nur ist (möglicherweise durch die Anlage der Untersuchungen) der Nachweis für einen solchen Zusammenhang noch nicht eindeutig erbracht worden. Auswirkungen des Fragebogenumfangs auf andere Variablen (insbesondere die Qualität der Antworten) und für andersartige Befragungstechniken (z.B. persönliche Befragung) wurden erst gar nicht untersucht.

Solche Untersuchungen müßten berücksichtigen, daß der Fragebogenumfang drei (nicht unabhängige, aber unterscheidbare) Aspekte, die Item-Anzahl, die Seitenzahl und die Bearbeitungsdauer, aufweist und daß der von der Vp erlebte Umfang nicht notwendig mit dem ‚objektiven‘ Umfang identisch sein muß (Richter 1969 und Erdos 1970 betonen z.B. die Wichtigkeit der Gliederung von Fragenserien; vgl. auch die Ausführungen über die äußere Gestaltung des Fragebogens, 5.). Besondere Schwierigkeiten für die Untersuchung der Auswirkungen des Fragebogenumfangs ergeben sich einmal aus den zu erwartenden Interaktionen mit anderen Merkmalen, dann aber auch aus der Tatsache, daß der Fragebogenumfang nicht ohne gleichzeitige Veränderung entweder des Fragebogeninhalts (bei einer Vermehrung der Zahl von Fragen) oder der Fragebogengestaltung (bei einer Verteilung der Fragen auf mehrere Seiten) vergrößert werden kann. Vielleicht liegt in diesen methodischen Schwierigkeiten eine Erklärung für den bemerkenswerten Mangel an empirischen Untersuchungen zur Rolle des Fragebogenumfangs.

5. Äußere Gestaltung (Layout) des Fragebogens

Fragen der typographischen und farblichen Gestaltung und des Layouts von Fragebogen sind kaum empirisch untersucht worden. Entsprechend gibt es sowohl was schriftlich zu bearbeitende Fragebogen (vgl. Hartley et al. 1977), als auch was Interviewer-Fragebogen für mündliche Befragungen betrifft (vgl. Haase 1978) wenig gesicherte Erkenntnisse. Andererseits berichtet Gray (1975) für unpersönlich-schriftliche (postalische) Befragungen durch Verbesserung der graphischen Gestaltung des Fragebogens im Vergleich zu einer maschinenschriftlichen ersten Version Rücklaufsteigerungen von ca. 30%, Ver-

minderungen der Bearbeitungszeiten von ca. 40 auf 20 Minuten und der Ablochzeiten um etwa 1/3. Rücklaufsteigerungen um immerhin noch 8% durch veränderte graphische Gestaltung fand auch Richter (1969). Es scheint also, daß gerade in der äußeren Aufmachung von Fragebogen erhebliche Möglichkeiten für eine Optimierung unter dem Kriterium der Qualität der Antworten und/oder unter ökonomischen Aspekten liegen.

Die *Verwendung von Farben* kann innerhalb eines Fragebogens unterschiedlichen Zielen dienen. Bei unpersönlich-schriftlichen (postalischen) Befragungen wird mitunter versucht, durch Wahl einer ansprechenden Papierfarbe den Rücklauf günstig zu beeinflussen, zumal damit keine ins Gewicht fallenden zusätzlichen Kosten verbunden sind. Sharma & Singh (1967) konnten - allerdings bei hochmotivierten und akademisch gebildeten Vpn mit einem Gesamtrücklauf von 87,7% - keinerlei Einfluß der Papierfarbe (weiß, rosa, gelb) auf den Rücklauf feststellen, ebensowenig gelang dies Gullahorn & Gullahorn (1963) für die Farben weiß und grün. Das bedeutet natürlich nicht, daß für andere Farben, in anderen Populationen und bei Fragebogen mit anderen formalen und inhaltlichen Merkmalen solche Einflüsse ebenfalls ausgeschlossen wären. Wegen der dadurch entstehenden Ähnlichkeit mit Werbedrucksachen warnt Erdos (1970) vor mehrfarbigen Fragebogen für unpersönlich-schriftliche (postalische) Befragungen.

Mit Vorteil läßt sich Farbe zur Kennzeichnung von Gabelungen und Verzweigungen in Fragebogen im Interesse einer besseren Handhabbarkeit durch den Interviewer einsetzen. Derartigen Farbkodierungen sind in der Praxis allerdings durch die hohen Kosten mehrfarbigen Druckes enge Grenzen gesetzt (Noelle 1963).

Ausgiebiger Gebrauch wird vom Medium Farbe bei Listen- und Kartenvorlagen gemacht, einmal im Interesse der Abwechslung für den Befragten (vgl. 4.2), aber auch zum Zwecke besserer Unterscheidbarkeit und eindeutiger Zuordnung zu den betreffenden Fragen für den Interviewer. Unproblematisch ist dies jedoch nur, wenn man davon ausgehen kann, daß die verwendete Farbe die Verteilung der Antworten (gewählten Karten) nicht beeinflusst. Ring (1969) versuchte dies für rote und graue Kartenvorlagen zu klären. Er kam zu dem Ergebnis, daß weder Zahl noch Art der Antworten durch die Hintergrundfarbe der Kartenvorlage beeinflusst wurden (die vereinzelt und unsystematisch aufgetretenen Antwortunterschiede in Abhängigkeit von der Farbe des Kartensatzes lassen sich als Produkte des Zufalls betrachten).

Obwohl es für die Motivation der Befragten vorteilhaft sein dürfte, wenn der Fragebogen durch Bedrucken der Vorder- und Rückseiten kurz erscheint (Erdos 1970), ist von diesem Vorgehen abzuraten: Fragen auf der Rückseite werden zu häufig übersehen (Kirschhofer-Bozenhardt & Kaplitza 1975).

Im Hinblick auf das Layout im engeren Sinne fordern Karmasin & Karmasin (1977) bei Fragebogen für mündliche Befragungen vor allem eine deutliche optische Unterscheidung (z.B. durch verschiedene Schrifttypen, Umrahmungen u.ä.) zwischen Anweisungen an den Interviewer, eigentlichem Fragentext und Antwortvorgaben. Nach Richter (1969) sollte bei schriftlich zu bearbeitenden Fragebogen (besonders im Falle unpersönlicher Befragung) auf Seiten der Vp der Eindruck vermieden werden, es handele sich um lange Fragenserien oder um viele einzelne Fragen, damit Ermüdungs-, Sättigungs- und Monotonieerlebnisse bei der Beantwortung möglichst gering gehalten werden können. Dazu schlägt er vor, einerseits der Einzelfrage nicht zuviel optisches Gewicht zu geben, sondern sie in einen Fragenblock einzugliedern, andererseits aber diese Fragenblöcke auch nicht zu umfangreich zu gestalten und sie durch Überschriften etc. voneinander abzuheben. Eine Beschreibung und Diskussion verschiedener Schrifttypen, Typengrößen, Satz- und Drucktechniken findet sich z.B. bei Erdos (1970), Gray (1975) und Wright & Barnard (1975).

Bei geschlossenen Fragen sind Kästchen oder Kreise vorzusehen, die der Vp anzeigen, wo sie ihre Markierung anzubringen hat. Handelt es sich um Fragenserien, sollten diese Kästchen bzw. Kreise eine klare graphische Anordnung (z.B. in einer Reihe untereinander) erhalten (Richter 1969) und zur Vermeidung von Verwechslungen nicht zu weit vom Text der Antworten entfernt sein (Wright & Barnard 1975). Richter (1969) fordert darüber hinaus, bei der Wahl der Größe für die Kästchen bzw. Kreise auf Besonderheiten der jeweiligen Zielpopulation Rücksicht zu nehmen. So seien bei älteren Menschen größere Kästchen bzw. Kreise erforderlich, aber auch z.B. Architekten unterschieden sich von z.B. Elektroingenieuren erheblich in der Größe der Kreuze, was bei ersteren im Interesse der Unmißverständlichkeit der Markierungen größere Kästchen bzw. Kreise erforderlich mache.

Hartley et al. (1977) untersuchten den Einfluß der Reihenfolge und genauen Anordnung von Antworttext, Kästchen und Codeziffern experimentell und stellten für die vier von ihnen verwendeten Varianten keine Auswirkungen auf das Antwortverhalten fest. Gewisse Unterschiede ergaben sich beim Zeitbedarf für die Erstellung des Fragebogenentwurfs, bei den Kosten für den Drucksatz und bei den Ablockkosten.

Zur Veranschaulichung und Verdeutlichung von Situationen bzw. Zusammenhängen, für die Beurteilungen oder Bewertungen erfragt werden sollen, z.T. aber auch im Interesse des Abwechslungsreichtums (vgl. 4.2), wird die Verwendung bildlicher Vorlagen empfohlen (z.B. Noelle 1963). Karmasin & Karmasin (1977) stellen jedoch fest, daß dabei mit subtilen, im einzelnen überwiegend nicht bekannten Einflüssen auf die Antworten zu rechnen sei. So werde z.B. ein Mann mit Hut innerhalb einer solchen bildlichen Vorlage als konservativer und besser situiert eingeschätzt als ein Mann ohne Hut, eine Hausfrau

mit sehr langem Haar gelte als weniger kompetent im Vergleich zu einer Frau mit kurzem Haar. In Abhängigkeit davon, wie bestimmte Rollenträger bzw. die Vertreter bestimmter Meinungen in den bildlichen Vorlagen dargestellt werden, sind dadurch Einflüsse auf die Antworten zu erwarten. Ring (1975) konnte Veränderungen in der Wahlhäufigkeit für bestimmte Statements in einer Größenordnung von 5% - 10% nachweisen, wenn die Zuordnung der Statements zu (stilisierten) Personen in einer bildlichen Vorlage vertauscht wurde. Nach Anlage dieser Untersuchung kann allerdings nicht entschieden werden, ob es sich dabei um Positioneffekte oder Einflüsse der Darstellungsweise der Personen handelt.

Bei der Erarbeitung des Fragebogen-Layouts muß auch festgelegt werden, wieviel *Antwortraum* bei offenen Fragen für die Eintragung der Antworten vorgesehen werden soll. Payne (1951) und Goode & Hatt (1972) berichten - gestützt auf entsprechende Erfahrungen - einen Anstieg der Antwortlänge mit Vergrößerung des für die Antworten vorgegebenen Raumes. Dies gelte einmal für schriftliche Befragungen, bei denen die Vp sieht, wieviel von ihr als Antwort erwartet wird, zum anderen aber auch für mündliche Befragungen, wobei ungeklärt sei, ob der Interviewer die Antworten ausführlicher protokolliert oder die Vp z.B. durch stärkeres Insistieren des Interviewers tatsächlich ausführlicher antwortet. Diese Frage griff Haase (1978) auf. Er ließ die Antworten der Vpn auf Tonband aufnehmen und stellte fest, daß - gemessen an der Zahl der Wörter - die Antworten bei Vergrößerung des für die Eintragung vorgesehenen Raumes tatsächlich länger wurden. Vom Antwortinhalt (Zahl der enthaltenen Antwortkategorien) her war die Ausführlichkeit der Antworten jedoch nicht unterschiedlich. Außerdem bestand eine Abhängigkeit vom Frageninhalt: Ein Anstieg der Antwortlänge (Wortanzahl) durch Vergrößerung des Antwortraumes konnte nur für Fragen nach Merkmalen einer kurzzeitig dargebotenen Anzeige und nach 'Gefühlen', die die Vpn mit dieser Anzeige verbinden, nicht aber für eine Frage nach bekannten Markenamen für ein bestimmtes Produkt aufgezeigt werden.

Bei schriftlichen Befragungen fand Tränkle (1974) Hinweise darauf, daß Antworten auch auf inhaltlicher Ebene (Zahl enthaltener Kategorien) ausführlicher waren, wenn 8 statt nur 3 Zeilen für die Eintragung der Antwort vorgesehen waren. Einflüsse des für Antworten vorgesehenen Raumes auf Antworten zu offenen Fragen scheinen also tatsächlich zu existieren, allerdings nur für bestimmte Frageninhalte, für einen bestimmten Variationsbereich des Antwortraumes (Haase 1978) und möglicherweise eher für die Form als für den Inhalt der Antworten.

Eine neuere, erst durch Einsatz von Textverarbeitungsanlagen realisierbar gewordene Entwicklung im Bereich der Fragebogengestaltung ist die *Individualisierung und Personalisierung* von Fragebogen. Für unpersönlich-schriftliche (postalische) Befragungen berichtet Perreault (1975) günstige Einflüsse auf den

Rücklauf, wenn der Fragebogen (scheinbar) individuell maschinenschriftlich erstellt, evtl. mit Namen und Adresse der Vp und mit dem Hinweis versehen ist, daß er in dieser Form nur an sie verschickt worden sei. Eingehender wird der Einsatz der Personalisierung im Interesse der Rücklaufsteigerung bei Erdos (1970) behandelt. In Abhängigkeit vom jeweiligen Befragungsgegenstand ist allerdings auch zu bedenken, daß die Personalisierung, wenn sie ihr Ziel erreicht, die extremste Form der Nicht-Anonymität und insofern ein ‚zweischneidiges Schwert‘ (Linsky 1975) ist.

6. Weitere Aspekte für die Konstruktion von Fragebogen

6.1 Anonymität des Befragten und Vertraulichkeit der Antworten

Als ‚anonym‘ wird eine Befragung dann bezeichnet, wenn es prinzipiell nicht möglich ist, ausgehend vom Fragebogen den Befragten zu identifizieren. Ist eine Befragung nicht anonym, aber vertraulich, so ist die Identität des Befragten zwar bekannt, wird aber gegenüber Dritten geheimgehalten (Dickson et al. 1977). Während die Zusicherung der Vertraulichkeit im Zusammenhang mit einer Befragung fast eine Selbstverständlichkeit zu sein scheint, wird die Notwendigkeit der Anonymität unterschiedlich beurteilt. Für persönlich mündliche Befragungen wird sie höchstens für einzelne Fragen angestrebt (vgl. z.B. die in 2.3 erwähnte ‚Urnentechnik‘), obschon sie objektiv auch für das ganze Interview dadurch gewährleistet werden könnte, daß ein Interviewer zahlreiche Interviews durchführt und auf die Kennzeichnung der Fragebogen verzichtet wird. Die für das Antwortverhalten der Vp einzig maßgebliche *erlebte* Anonymität wird allerdings für persönlich-mündliche Befragungen kaum zu erreichen sein. Demzufolge beschränkt sich die Diskussion auch auf schriftliche, insbesondere unpersönlich-schriftliche, z.B. postalische Befragungen.

Einige Autoren berichten für diese Befragungsform bedeutsame Antwortunterschiede in Abhängigkeit von der Anonymität bzw. Nicht-Anonymität der Befragten. Knudsen et al. (1967) fanden, daß in persönlichen Befragungen restriktivere Normen betreffend den vorehelichen Geschlechtsverkehr vertreten wurden als in unpersönlichen und anonymen Befragungen. Auch Fuller (1974) und Bradburn & Sudman (1979) berichten Antwortverzerrungen nach Maßgabe sozialer Erwünschtheit bei nicht-anonymer im Vergleich zu anonymer Befragung. Als Hinweise in dieser Richtung könnten auch die Ergebnisse von Taietz (1972) angesehen werden, der bei der Befragung älterer Menschen nach ihren Lebensverhältnissen erhebliche Verschiebungen in den Antworten dann erhielt, wenn eine dritte Person beim Interview anwesend war (vgl. auch Bradburn & Sudman 1979).

Verzerrungen in genau entgegengesetzte Richtung fanden Epperson & Peck (1977) in einer Untersuchung zur Evaluation von Driver-Improvement-Programmen (vgl. dazu Spoerer 1979). Hier fanden sich signifikant mehr negative Kommentare der Teilnehmer, wenn die Befragung nicht anonym durchgeführt wurde. Insgesamt sind die Nachweise bedeutsamer Antwortunterschiede in Abhängigkeit von der Anonymität jedoch eher spärlich. Kepes & True (1967) und auch Fuller (1974) kommen bei der Sichtung empirischer Befunde zu dem Ergebnis, daß der Einfluß der Nicht-Anonymität auf Antworten eher nur befürchtet als real sei. Die erstgenannten Autoren sehen ihn - außer in einigen ziemlich speziellen Situationen - vor allem dann, wenn die Vp eigens und explizit auf die Namentlichkeit hingewiesen wird, wozu in der Regel aber keine Notwendigkeit besteht. Daß den Vpn in vielen Fällen das Nicht-Vorliegen von Anonymität bzw. Vertraulichkeit gar nicht bewußt ist und der Einfluß von Anonymität bzw. Vertraulichkeit zutreffend nur nach entsprechendem explizitem Hinweis abgeschätzt werden kann, betonen auch Futrell & Swan (1977). Sie fanden zwischen anonymer und nicht-anonymer, aber vertraulicher postalischer Befragung keinerlei Unterschiede und sehen in der Anonymität keine Vorteile, wenn Untersucher und Auftraggeber nicht identisch sind und den Vpn Vertraulichkeit zugesichert werden kann, Butler (1973) ließ Fragen, die sich unter anderem auf Drogenkonsum bezogen, von Experten hinsichtlich ihrer Anfälligkeit für Antwortverzerrungen bei Verwendung in nicht-anonymer Befragung skalieren, bevor er sie Kadetten einer Militärakademie teils anonym, teils nicht-anonym zur Beantwortung vorlegte. Er fand bei Fragen, die die Experten als ‚unempfindlich‘ eingestuft hatten, erwartungsgemäß keine Beantwortungsunterschiede, wider Erwarten unterschieden sich anonyme und nicht-anonyme Antworten aber auch bei den als ‚empfindlich‘ klassifizierten Fragen nicht. Neben anderen möglichen Erklärungen könnte auch hier die den nicht-anonym antwortenden Vpn gegebene Vertraulichkeitszusage zur Vermeidung von Verzerrungen ausgereicht haben. Keine Unterschiede zwischen anonymen und nicht-anonymen Antworten von Lehrern auf Fragen zur Beurteilung der Notwendigkeit gewerkschaftlicher Organisation und der eigenen Streikbereitschaft fand auch Wildman (1977). Andererseits berichtet er aber, daß im Rahmen dieser postalischen Befragung 12% der nicht-anonymen Vpn die Identifikationsnummern auf ihren Antwortbogen vor der Rücksendung unkenntlich gemacht hatten, was sich wohl nur auf ein Bedürfnis nach Anonymität zurückführen läßt.

Einflüsse der Anonymität auf den Rücklauf in unpersönlich-schriftlichen (postalischen) Befragungen sind nach Richter (1969) zwar je nach Zielpopulation unterschiedlich, insgesamt aber nicht ‚durchschlagend‘. Bei Bradburn & Sudman (1979) fanden sich geringe, bei Wildman (1977) keinerlei Unterschiede im Rücklauf zwischen anonym und nicht-anonym befragten Vpn. Fuller (1974) berichtet sogar - abweichend von der landläufigen Erwartung - bei Nicht-Anonymität einen höheren Rücklauf (evtl. ein Personalisierungseffekt, vgl. 5.).

Während bei mündlichen und persönlichen schriftlichen Befragungen auch andere Gründe (z.B. die Notwendigkeit der Kontrolle von Interviewern) für den Verzicht auf Anonymität in Betracht kommen, ist das Interesse an der Identifizierbarkeit der Befragten in unpersönlich-schriftlichen (postalischen) Befragungen weitgehend in dem Wunsch nach Kontrollierbarkeit und gezielter Beeinflussung des Rücklaufs begründet. Einerseits sprechen Kostenerwägungen, andererseits aber auch die Gefahr von Doppel-Beantwortungen dagegen, die Fragebogen mehrfach an alle Vpn zu verschicken. Die Möglichkeit einer gezielten Erinnerung der Nicht-Beantworter besteht aber natürlich nur bei Identifizierbarkeit der Rückläufe. Um dennoch die vermuteten Vorteile der Anonymität nutzen zu können, hat in den USA die unsichtbare Kennzeichnung der Fragebogen eine weite Verbreitung gefunden, eine Praxis, die aus ethischen und juristischen Gründen zweifellos abzulehnen ist (vgl. Dickson et al. 1977). Alternativen, die eine Kontrolle der Rückläufe trotz strikter Anonymität gestatten, beschreibt z.B. Wieken (1974). So kann man dem Fragebogen eine mit der Adresse der Vp als Absender versehene frankierte Postkarte beifügen und die Vp bitten, diese gleichzeitig mit, aber getrennt von dem nicht gekennzeichneten Fragebogen zurückzuschicken, damit der Untersucher weiß, daß, aber nicht was sie geantwortet hat (Linsky 1975).

6.2 Spezielle Probleme bei unpersönlich-schriftlichen Befragungen

Unpersönlich-schriftliche Befragungen sind dadurch gekennzeichnet, daß ein Fragebogen in Abwesenheit des Interviewers bearbeitet wird. Der Fragebogen kann dem Befragten persönlich übergeben oder z.B. mit der Post zugeschickt worden sein. Diese letztgenannte Form der Befragung, die sogenannte postalische Befragung, erfreut sich aus mehreren Gründen vergleichsweise großer Beliebtheit, deren wichtigster die relativ geringen Kosten sein dürften (Stroschein 1965, Richter 1969, Goode & Hatt 1972, Wieken 1974). Allerdings sind zum Zwecke einigermaßen akzeptabler Stichprobenausschöpfungen fast immer mehrere Befragungswellen oder Erinnerungsschreiben notwendig, so daß den niedrigen Kosten ein vergleichsweise hoher Zeitbedarf (selten weniger als 6-8 Wochen, vgl. z.B. Buchner 1968) für die Datenerhebung gegenübersteht. Einerseits eignet sich die Methode damit für die Gewinnung aktueller Daten prinzipiell nicht, andererseits ist sie besonders anfällig gegenüber unvorhergesehenen, während der langen Erhebungsphase wirksam werdenden Einflüssen (z.B. Veröffentlichungen zum Thema).

Sachliche Vorteile der postalischen Befragung liegen für bestimmte Fragestellungen (wie bei allen schriftlichen Befragungsformen) in ihrem unpersönlichen und gegebenenfalls anonymen Charakter, der Antwortverzerrungen etwa nach Maßgabe sozialer Erwünschtheit weniger wahrscheinlich macht (Tränkle

1974). So berichtet etwa Friedrich (1970), daß Antworten in schriftlichen Befragungen weniger stark gesellschaftlichen Normen entsprechen als in mündlichen Interviews. Metzner & Mann (1952) erhielten für die Zufriedenheit von Arbeitnehmern mit ihren Vorgesetzten in schriftlichen verglichen mit mündlichen Befragungen ungünstigere Antworten. Auch Linsky (1975) sieht in der geringeren sozialen Kontrolle, der das Antwortverhalten unterliegt, einen wesentlichen Vorteil der schriftlichen Befragung.

Weitere Vorteile können - je nach Fragestellung und Zielpopulation - auch darin liegen, daß zur Beantwortung der Fragen Unterlagen herangezogen werden und an der Beantwortung mehrere Personen mitwirken können (Linsky 1975). In den meisten Fällen aber wird die bei der unpersönlichen Befragung prinzipiell fehlende Möglichkeit der Kontrolle von Beantwortungsperson, Beantwortungssituation, Beantwortungszeitpunkt und Reihenfolge der Beantwortung der Fragen als Nachteil betrachtet werden müssen.

Ein noch gravierenderer Nachteil der unpersönlichen, meist postalisch durchgeführten Befragung liegt in den relativ hohen Anforderungen, die sie an die Befragten stellt und die mindestens für den Teil der Bevölkerung, den Scheuch (1973) den ‚funktionellen Analphabeten‘ zuordnet, zu hoch sein dürften. Auch Kreutz & Titscher (1974, 60) stellen fest, daß „... in weiten Kreisen der Bevölkerung Angst vor Rechtschreibfehlern und Schwierigkeiten bei der schriftlichen Formulierung bestehen ...“, und halten mindestens offene Fragen in unpersönlich-schriftlichen Befragungen dann für kontraindiziert, wenn die Zielpopulation nicht z.B. durch Bildung bzw. Beruf sprachlich besonders geübt ist.

Da der Rücklauf in postalischen Befragungen positiv mit der sozialen Schicht, dem Bildungs-, Berufs- und Einkommensniveau korreliert (z.B. Richter 1967, 1969, Goode & Hatt 1972, Wieken 1974, Binder et al. 1979), ist bei inhomogenen Stichproben (z.B. Bevölkerungsstichproben) mit systematischen Verzerrungen dadurch zu rechnen, daß Angehörige unterer sozialer Schichten mit niedrigem Bildungs-, Berufs- und Einkommensniveau in der Gruppe der antwortenden Vpn unterrepräsentiert sind. Unpersönlich-schriftliche (postalische) Befragungen sollten deshalb nur für homogene lese- und schreibgewandte Populationen in Betracht gezogen werden. Nach Kish & Barnes (1973) eignet sich die postalische Befragung außerdem nicht für Befragungsinhalte, die in der Zielpopulation kontrovers eingeschätzt werden: Der Rücklauf erwies sich als umgekehrt proportional der Strittigkeit der Inhalte.

Auch wenn z.B. Mc Donagh & Rosenblum (1965) in einer mündlichen Nachbefragung von Antwortern und Nicht-Antwortern einer vorangegangenen postalischen Befragung keinerlei Beantwortungsunterschiede feststellen konnten und deshalb annehmen, das Problem der Irrepräsentativität der Antworter für die gesamte Population werde üblicherweise überschätzt, läßt sich die

Gefahr der Irrepräsentativität natürlich nie prinzipiell ausschließen, Binder et al. (1979) etwa fanden beträchtliche Unterschiede zwischen Antwortern und Nicht-Antworthen in demographischen und Persönlichkeitsmerkmalen. Es ist deshalb von besonderer Wichtigkeit, durch geeignete Anlage der Untersuchung (z.B. Vorkontakten der Vpn; Gestaltung von Begleitschreiben, Fragebogen; Rückumschlag; mehrfache Erinnerungsschreiben) für eine möglichst hohe Stichprobenausschöpfung zu sorgen. Entsprechende Hinweise und empirische Ergebnisse finden sich z.B. bei Richter (1969), Alutto (1970), Erdos (1970), Hendrick et al. (1972), Kish & Barnes (1973), Wieken (1974) und Sieber (1979). Linsky (1975) hat eine ausführliche Zusammenstellung empirischer Befunde zur Frage der Beeinflußbarkeit des Rücklaufs erarbeitet, die u.a. erkennen läßt, daß es kaum Befunde betreffend den Einfluß der Gestaltung des Fragebogens auf den Rücklauf gibt (vgl. 5.).

Neben und evtl. zusätzlich zu dem Bemühen um einen möglichst hohen Rücklauf werden gelegentlich auch Korrekturen der Ergebnisse zum Ausgleich etwa bestehender Irrepräsentativität vorgenommen. Solche Korrekturen beruhen natürlich auf Annahmen betreffend das potentielle Antwortverhalten der Nicht-Antworthenden. Meist gehen sie davon aus, daß die Vpn, die zuletzt geantwortet haben, „Beinahe-Nicht-Antworter“ sind, und verwenden deren Antwortverhalten zur Schätzung des Verhaltens der Nicht-Antworter. Evtl. wird auch versucht, einen Trend, der sich im Antwortverhalten von den frühesten zu den letzten Rückläufen hin zeigt, auf die Nicht-Antworter zu extrapolieren. Überlegungen zur Repräsentativitätskorrektur finden sich u.a. bei Richter (1967, 1969), Buchner (1968), Erdos (1970) und Wieken (1974).

6.3 Erprobung und Überarbeitung des Fragebogenentwurfs

Itemanalysen, Itemselektion und Konstruktion abgeleiteter Variabler (z.B. Summierung von Antworten) erfolgen bei diagnostischen Fragebogen wie bei jedem Test auf der Grundlage eines bestimmten Meßmodells und Validitätskonzeptes (vgl. z.B. Lienert 1969, Fischer 1974, Wottawa 1980). Darüber hinaus sind jedoch auch fragebogenspezifische Gütekriterien (Verstehbarkeit, Ambiguität, soziale Erwünschtheit der Items) und Eigentümlichkeiten (z.B. veränderte Bedeutung der Itemschwierigkeit) zu beachten (Janke 1973). Hier auf soll an dieser Stelle nicht näher eingegangen werden.

Bei Fragebogen mit sozialwissenschaftlicher bzw. demoskopischer Zielsetzung wird die Notwendigkeit des Pretests (Karmasin & Karmasin 1977 fordern dafür sogar 100 Vpn) und der Revision des Fragebogens zwar allgemein anerkannt bzw. hervorgehoben, doch charakterisieren Cannell et al. (1977, 27) die in der Praxis übliche Vorgehensweise (in der Übersetzung des Verfassers) folgendermaßen:

„Normalerweise erstellt man den Fragebogenentwurf am Schreibtisch und schickt dann eine Gruppe von Interviewern damit ins Feld. Danach gibt es eine Konferenz (oder eine Serie von Konferenzen), auf der Forscher und Interviewer den Fragebogen diskutieren. Man hört dabei Aussagen wie „ . . . diese Frage scheint gut zu funktionieren . . . “ oder der Interviewer sagt: „ . . . Ich glaube nicht, daß die Befragten diese Frage wirklich verstanden haben . . . “. Auf der Grundlage derart subjektiver Bewertungen werden Fragebogen üblicherweise entwickelt.“ Selbstverständlich können die Erfahrungen der Interviewer einen wichtigen Beitrag zur Revision des Fragebogens leisten, nur sollte es nicht deren einzige Grundlage sein. Guski et al. (1978) beschreiben beispielhaft die Konstruktion eines sozialwissenschaftlichen Fragebogens zur Erfassung von Auswirkungen des Umweltlärms. Ausgehend von den Ergebnissen einer Vorstudie (Explorationen mit 30 Vpn), von einer Inhaltsanalyse der Beschwerden über Lärmbelästigung, die bei Behörden eingegangen waren, und von bereits existierenden Fragebogen zum Thema wurde ein Fragebogenentwurf erstellt und einem Pretest an 40 Vpn unterworfen. Statistische Itemanalysen (dazu können wie bei der Konstruktion diagnostischer Fragebogen u.a. Verteilungs-, Schwierigkeits-, Trennschärfe- und Interkorrelationsanalysen gehören; vgl. Berk & Griesemer 1976) und Interviewererfahrungen bildeten die Grundlage einer Revision des Fragebogens für die Hauptuntersuchung an über 600 Vpn. Damit waren die methodologischen Bemühungen um den Fragebogen allerdings nicht abgeschlossen, vielmehr wurden die Definitionen abgeleiteter Variabler (z.B. die Summierungen von Reaktionen auf verschiedene Items) mit den Daten der Hauptuntersuchung (zur Prüfung der Stabilität gegenüber einer Variation der Stichprobe meist getrennt für zwei Zufallshälften der Stichprobe) jeweils empirisch abgesichert, vor allem mittels Cluster- und Faktorenanalysen. Wegen der hierfür erforderlichen hohen Vpn-Zahl wäre es unrealistisch, solche Analysen schon im Stadium des Pretests zu verlangen, vielmehr wird man die Fragebogenkonstruktion und -Überprüfung als einen Prozeß auffassen müssen, der nie abgeschlossen, sondern höchstens abgebrochen werden kann.

Die Aufgaben, die der Pretest erfüllen kann und erfüllen muß, nämlich die Überprüfung von Fragenformulierungen, Fragebogaufbau und -gestaltung, werden um so wichtiger, je weniger Kompensationsmöglichkeiten für Mängel des Fragebogens in der Befragungssituation selbst vorhanden sind. Von besonderer Bedeutung ist die Erprobung des Fragebogens demnach für alle nicht-persönlichen Befragungen, also besonders für die postalische Befragung. Richter (1969) spricht in diesem Zusammenhang von der Notwendigkeit, den Fragebogen im „Putzfrauentest“, d.h. durch Anwendung bei den sprachlich und intellektuell am wenigsten differenzierten Vpn der Zielpopulation zu erproben.

Für diagnostische Fragebogen mit längerer Lebensdauer sind - wie für jeden Test - kontinuierliche Kontrolluntersuchungen erforderlich (Lennertz 1973).

So weist z.B. Strong (1962) auf die Notwendigkeit hin, veraltende Inhalte von Items (Persönlichkeiten, Buch- und Filmtitel etc.) entweder grundsätzlich zu meiden oder aber häufigere Revisionen und Aktualisierungen der Fragebogen durchzuführen. Ash & Edgell (1975) demonstrierten die Nichtübereinstimmung des sprachlichen Niveaus des Position-Analysis-Questionnaire (PAQ) von Mc Cormick (vgl. Mc Cormick et al. 1965) mit demjenigen der tatsächlichen Anwender und machten deutlich, daß auch Änderungen der Zielpopulation in Rechnung zu stellen sind.

Die Kontrolle des Fragebogenentwurfs muß sich sodann natürlich auf Fragen der Reliabilität und Validität (bzw. Generalisierbarkeit im Sinne von Cronbach et al. 1972) erstrecken. Wie für alle Tests so ist auch für diagnostische Fragebogen unbestritten, daß Aussagen über ihre Güte nur in bezug auf ein bestimmtes Meßmodell, auf ein bestimmtes Validitätskonzept und evtl. auf eine bestimmte Population möglich bzw. sinnvoll sind. Ebenso ist es für sozialwissenschaftliche bzw. demoskopische Anwendungen abwegig, die Qualität von Fragebogenerhebung bzw. Interview allgemein feststellen zu wollen, wie dies etwa Friedrich (1963, 1966) und Förster (1967) für schriftliche und Fisseni (1974) für mündliche Befragungen zu tun versuchen (vgl. auch Sieber 1979). Aussagen sind auch hier nur möglich für die Methode bezogen auf einen Gegenstand und eine bestimmte Population. Bei mündlichen und persönlich-schriftlichen Befragungen sind außer dem Fragebogen die Interviewer, bei unpersönlich-schriftlichen Befragungen die Techniken und der Grad der Stichprobenausschöpfung zentrale Bestandteile der Methode. Nachweise hoher Objektivität, Reliabilität und Validität stellen hier bestenfalls Existenzbeweise dar.

7. Zukünftige Entwicklung im Bereich der Fragebogenkonstruktion

Um je nach gewähltem Validitätskonzept und Meßmodell (vgl. 1.1.2) die Konzeption einer Frage auf empirisch gesicherter Basis entwickeln zu können, ist es erforderlich, das Wissen über den Beantwortungsprozeß und die Determinanten der Antwort (vgl. 1.2) zu erweitern. Erhebliche Wissenslücken bestehen sodann im Bereich der sprachlichen Formulierung der Frage (3.2), der Fragenreihenfolge (4.) und vor allem der Auswirkungen der äußeren Gestaltung des Fragebogens (5.) auf das Beantwortungsverhalten.

Für *demoskopische (sozialwissenschaftliche)* Fragebogen zeichnet sich durch die leichtere Verfügbarkeit elektronischer Datenverarbeitungsanlagen ein Verschwinden des für alle Vpn einheitlichen Fragebogens zugunsten einer größeren Zahl von Fragebogenvarianten mit variiertem Reihenfolge der Fragen, variierten Frageformulierungen, variiertem äußerer Gestaltung bis hin zum indivi-

dualisierten und möglicherweise personalisierten Fragebogen ab (Perreault 1975). Dabei ist es durchaus auch möglich, unter Verwendung von Vorinformationen über den Befragten eine ganz spezielle Fragenzusammenstellung zu konzipieren und damit Filterungen und Verzweigungen, wie sie in traditionellen Fragebogen erforderlich sind, entbehrlich zu machen, was besonders für unpersönlich-schriftliche (postalische) Befragungen die möglichen Befragungsinhalte erheblich ausweiten dürfte. Darüber hinaus lassen sich unter Nutzung elektronischer Datenverarbeitungsanlagen Fragenpools aufbauen, die eine rasche Ad-hoc-Konstruktion von Fragebogen für bestimmte Anwendungen ermöglichen (Doyle & Wattawa 1977).

Inwieweit neue elektronische Medien, wie Videotext und Telekommunikation, auch die Durchführung von Befragungen nachhaltig verändern werden, ist derzeit nicht abzuschätzen. In Anlehnung an die Erfahrungen mit Telefon-Interviews ist jedoch zu vermuten, daß es das Bildschirm-Interview für bestimmte Untersuchungen geben wird, daß es die traditionellen Befragungsformen jedoch nicht wird verdrängen können.

Die statistische Auswertung von Fragebogendaten, die heute noch überwiegend einzelfragenorientiert erfolgt, wird sich zunehmend der angemesseneren multivariaten Analyse- und Testverfahren bedienen (vgl. Whitney & Feldt 1973).

Im Bereich *diagnostischer Fragebogen* werden sich die test- und meßtheoretischen Grundlagen weiterentwickeln. Dabei dürfte einerseits dem ordinalen Charakter von Fragebogendaten stärker Rechnung getragen, andererseits dürfen aber auch Versuche unternommen werden, die Datenqualität in Richtung auf metrische Eigenschaften zu verbessern. Dabei haben mehrkategoriale probabilistische Modelle gerade für Fragebogen große Bedeutung.

Erhebliche Möglichkeiten scheinen auch in der Anwendung der Methoden individualisierten (antwortabhängigen) Testens im Falle von Fragebogen zu liegen; Versuche in dieser Richtung beschreibt z.B. Hornke (1979). In gewisser Hinsicht handelt es sich dabei um die Realisierung der auch für demoskopische Fragebogen gebräuchlichen Techniken der Filterung und Verzweigung: Während in einem herkömmlichen diagnostischen Fragebogen jeder Proband alle Fragen zu bearbeiten hat, werden beim antwortabhängigen Test diejenigen Items nicht dargeboten, die zur (zuverlässigen) Schätzung des Ortes des Probanden auf der interessierenden Dimension nichts Wesentliches beitragen.

Außer zur Datengewinnung im Bereich sozialwissenschaftlicher Fragestellungen und zu diagnostischen Zwecken sind Fragebogen auch mit dem Ziel der Änderung von Einstellungen (Dillehay & Jernigan 1970) und mit therapeutischer Intention als Hilfsmittel bei der Selbsterfahrung (Hendrix 1978) eingesetzt worden. Es ist schwer abzuschätzen, ob sich diese Anwendungen bewäh-

ren und vermehren und ob sich weitere Einsatzmöglichkeiten eröffnen werden.

Umgekehrt dürften Fragebogen überall dort ihre Berechtigung verlieren, wo direktere und objektivere Methoden der Datengewinnung verfügbar werden. So können bestimmte Daten aus diagnostischen Fragebogen möglicherweise durch physiologische Messungen ersetzt werden: Statt nach Schlafqualität zu fragen, kann man sie u.U. dem EEG entnehmen. *Demoskopische bzw. sozialwissenschaftliche* Fragebogen dürften in den Bereichen entbehrlich werden, in denen vorhandene Dateien abgefragt werden können (z.B. muß der Führerscheinbesitz z.Z. noch durch Befragung erhoben werden, nach Aufbau einer entsprechenden Datei würde diese Notwendigkeit entfallen).

Für die im Bereich der Diagnostik wie der sozialwissenschaftlichen Datenerhebung wichtigen Beurteilungen und Bewertungen durch Personen sind zwar Alternativen zur hergebrachten Form des Fragebogens, nicht aber zur Methode der Befragung erkennbar.

Literatur

- Adams, J. S. 1956. An experiment on question and response bias. *Public Opinion Quarterly*, 20, 593-598.
- Alutto, J. A. 1970. Some dynamics of questionnaire completion and return among professional and managerial personnel. *Journal of Applied Psychology*, 54, 430-432.
- Anastasi, A. 1968. *Psychological testing*. London: Macmillan.
- Andersen, E. B. 1973. Conditional inference for multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 26, 31-44.
- Anger, H. 1969. Befragung und Erhebung. In: Graumann, C. F. (Hrsg.): *Handbuch der Psychologie*, Bd. 7: Sozialpsychologie, 1. Halbbd. Göttingen: Hogrefe.
- Ash, R. & Edgell, S. L. 1975. A note on the readability of the position analysis questionnaire (PAQ). *Journal of Applied Psychology*, 60, 775-776.
- Atteslander, P. 1971. *Methoden der empirischen Sozialforschung*. Berlin: De Gruyter.
- Barton, A. H. 1958. Asking the embarrassing question. *Public Opinion Quarterly*, 22, 67-68.
- Behrens, K. C. (Hrsg.) 1974. *Handbuch der Marktforschung*. Wiesbaden: Gabler.
- Belson, W. A. 1966. The effect of reversing the presentation order of verbal rating scales. *Journal of Advertising Research*, 6, 30-37.
- Berdie, D. R. 1973. Questionnaire length and response rate. *Journal of Applied Psychology*, 58, 278-280.

- Berg, I. A. 1967. Response set in personality assessment. Chicago: Aldine.
- Berk, R. A. & Griesemer, H. A. 1976. Iteman: An item analysis program for tests, questionnaires and scales. *Educational and Psychological Measurement*, 36, **189-191**.
- Binder, J., Sieber, M. & Angst, J. 1979. Verzerrungen bei postalischer Befragung: das Problem der Nichtantworter. *Zeitschrift für experimentelle und angewandte Psychologie*, 24, 53-71.
- Block, J. 1965. The challenge of response sets. New York: Appleton Century Crofts.
- Bradburn, N. M. & Mason, W. M. 1964. The effect of question order on response. *Journal of Marketing Research*, 1, 57-61.
- Bradburn, N. M. & Sudman, S. 1979. Improving interview method and questionnaire design. London: Jossey Bass.
- Buchner, D. 1968. Probleme und Antwortmuster bei postalischen Ärztebefragungen. *Der Marktforscher*, 7, 178-181.
- Burisch, M. 1976. Konstruktionsstrategien für multidimensionale Persönlichkeitsfragebögen. Hamburg: Phil. Diss.
- Butler, R. P. 1973. Effects of signed and unsigned questionnaires for both sensitive and nonsensitive items. *Journal of Applied Psychology*, 57, 348-349.
- Cahalan, D., Tamulonis, V. & Verner, H. W. 1947. Interviewer bias involved in certain types of opinion survey questions. *International Journal of Opinion and Attitude Research*, 1, 63-77.
- Campbell, D. T. & Fiske, D. W. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Cannell, C. F. & Kahn, R. L. 1968. Interviewing. In: Lindzey, G. & Aronson, E. (eds.): *Handbook of Social Psychology*, Vol. 2. Reading: Addison-Wesley.
- Cannell, C. F., Marquis, K. H. & Laurent, A. 1977. A summary of studies of interviewing methodology. Rockville: U.S. Department of Health, Education and Welfare.
- Carl, W. 1968. Eine Untersuchung zur Faktorenstruktur von Antworttendenzen bei Antwortskalen unterschiedlicher Stufenzahl. *Zeitschrift für experimentelle und angewandte Psychologie*, 15, 419-434.
- Cataldo, E. F., Johnson, R. M., Kellstedt, L. A. & Milbrath, L. W. 1970. Card sorting as a technique for survey interviewing. *Public Opinion Quarterly*, 34, 202-215.
- Cattell, R. B. 1974. How good is the modern questionnaire? General principles for evaluation. *Journal of Personality Assessment*, 38, 115-129.
- Cattell, R. B., Eber, W. H. & Tatsuoka, M. M. 1970. Handbook for sixteen personality factor questionnaire. Champaign: Institute for Personality and Ability Testing.
- Clauss, G. 1968. Zur Methodik von Schätzskalen in der empirischen Forschung. *Probleme und Ergebnisse der Psychologie*, 26, 7-53.
- Cliff, N. 1977. Further study of cognitive processing models for inventory response. *Applied Psychological Measurement*, 1, 41-49.

- Cliff, N., Bradley, P. & Girard, R. 1973. The investigation of cognitive models for inventory response. *Multivariate Behavioral Research*, 8, 407-425.
- Coan, R. W. 1964. Facts, factors and artifacts: The quest for psychological meaning. *Psychological Review*, 71, 123-140.
- Cohen, R. & Carl, W. 1964. Beantwortungsstereotypen (response sets) im Polaritätsprofil und ihre Beziehung zum Neurotizismus. *Diagnostica*, 10, 133-144.
- Cronbach, L. J. 1970. *Essentials of psychological testing*. New York: Harper & Row.
- Cronbach, L. J. & Meehl, P. E. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- Cronbach, L.J., Gleser, G. C., Nanda, H. & Rajaratnam, N. 1972. The dependability of behavioral measurement. New York: Wiley.
- Crutchfield, R. S. & Gordon, D. A. 1947. Variations in respondent interpretations of an opinion poll question. *International Journal of Opinion and Attitude Research*, **1, 1-12.**
- Damarin, F. 1970. A latent structure model for answering personal questions. *Psychological Bulletin*, 73, 23-40.
- Dickson, J. P., Casey, M., Wyckoff, D. & Wynd, W. 1977. Invisible coding of survey questionnaires. *Public Opinion Quarterly*, 41, 100-106.
- Dillehay, R. & Jernigan, L. R. 1970. The biased questionnaire as an instrument of opinion Change. *Journal of Personality and Social Psychology*, 15, 144-150.
- Doyle, K. C. & Wattawa, S. 1977. Programs for the construction and analysis of custom questionnaires and rating scales. *Educational and Psychological Measurement*, 37, 237-239.
- Edwards, A. L. 1957. *Techniques of attitude scale construction*. New York: Appleton Century Crofts.
- Edwards, A. L. 1970. *The measurement of personality traits by scales and inventories*. New York: Holt, Rinehart & Winston.
- Ehlers, T. 1973. Zur Effektivität der Kontrollen von Reaktionseinstellungen. In: Reinert, G. (Hrsg.): Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970. Göttingen: Hogrefe.
- Ellis, A. 1947. A comparison of the use of direct and indirect phrasing in personality questionnaires. *Psychological Monographs*, 61, whole No. 284.
- Epperson, W. V. & Peck, R. C. 1977. Questionnaires response bias as a function of respondent anonymity. *Accident Analysis and Prevention*, 9, 249-256.
- Erdos, P. L. 1970. *Professional mail surveys*. New York: McGraw Hill.
- Eysenck, H. J. 1953. Fragebogen als Meßmittel der Persönlichkeit. *Zeitschrift für experimentelle und angewandte Psychologie*, 1, 291-335.
- Eysenck, H. J. 1956. *Wege und Abwege der Psychologie*. Reinbek: Rowohlt.
- Falzhzik, A. M. & Carroll, S. J. 1971. Rate of return for closed and opened questions in a mail questionnaire survey of industrial Organisation. *Psychological Reports*, 29, **1121-1122.**

- Feger, H. 1974. Die Erfassung individueller Einstellungsstrukturen. *Zeitschrift für Sozialpsychologie*, 5, 242-254.
- Fischer, G. 1974. Einführung in die Theorie psychologischer Tests. Bern: Huber.
- Fiske, D. W. 1978. *Strategies for personality research*. San Francisco: Jossey Bass.
- Fisseni, H. J. 1974. Zur Zuverlässigkeit von Interviews. *Archiv für Psychologie*, 126, 71-84.
- Förster, P. 1967. Zu einigen methodischen Problemen der schriftlichen Befragung. *Jugendforschung*, 1/2, 39-67.
- Friedrich, W. 1963. Die Befragungsmethode: ein notwendiges Arbeitsmittel der marxistischen Jugendforschung. *Deutsche Zeitschrift für Philosophie*, 10, **1230-1247**.
- Friedrich, W. 1966. Zur Reliabilität von schriftlichen Befragungen. *Wissenschaftliche Zeitschrift der Karl-Marx-Universität Leipzig*, 15, 805-808.
- Friedrich, W. (Hrsg.) 1971. *Methoden der marxistisch-leninistischen Sozialforschung*. Berlin: VEB Deutscher Verlag der Wissenschaften.
- Friedrich, W. & Hennig, W. (Hrsg.) 1975. *Der sozialwissenschaftliche Forschungsprozeß*. Berlin: VEB Deutscher Verlag der Wissenschaften.
- Friedrichs, J. 1973. *Methoden empirischer Sozialforschung*. Reinbek: Rowohlt.
- Frisbie, B. & Sudman, S. 1968. The use of computers in coding free responses. *Public Opinion Quarterly*, 32, 216-232.
- Fürntratt, E. 1969. Antworttendenzen in Fragebogen 1: Bejahungs- und Varianztendenzen. *Psychologische Rundschau*, 20, 1-18.
- Fuller, C. 1974. Effect of anonymity on return rate and response bias in a mail survey. *Journal of Applied Psychology*, 59, 292-296.
- Futrell, C. M. & Swan, J. E. 1977. Anonymity and response by salespeople to a mail questionnaire. *Journal of Marketing Research*, 14, 611-616.
- Galtung, J. 1973. *Theory and methods of social research*. London: Allen & Unwin.
- Getzels, J. W. 1954. The question-answer process: a conceptualisation and some derived hypotheses for empirical examination. *Public Opinion Quarterly*, 18, 80-91.
- Goode, W. J. & Hatt, P. K. 1972. Die schriftliche Befragung. In: König, R. (Hrsg.): *Das Interview*. Köln: Kiepenheuer & Witsch.
- Gray, A. 1975. Questionnaire typography and production. *Applied Ergonomics*, 6, 81-89.
- Guilford, J. P. 1954. *Psychometric methods*. New York: McGraw Hill.
- Guilford, J. P. 1965. *Persönlichkeit*. Weinheim: Beltz.
- Gullahorn, J. E. & Gullahorn, J. T. 1963. An investigation of the effects of three factors on response to mail questionnaires. *Public Opinion Quarterly*, 27, **294-296**.
- Gulliksen, H. 1950. *Theory of mental tests*. New York: Wiley.

- Guski, R., Wichmann, U., Rohrmann, B. & Finke, H. O. 1978. Konstruktion und Anwendung eines Fragebogens zur sozialwissenschaftlichen Untersuchung der Auswirkungen von Umweltlärm. *Zeitschrift für Sozialpsychologie*, 9, 50-65.
- Haase, H. 1978. Zum Einfluß des Fragebogen-Layouts auf Befragungsergebnisse. In: Hartmann, K. D. & Koeppler, K. (Hrsg.): *Fortschritte der Marktpsychologie*, Bd. 1. Frankfurt: Fachbuchhandlung für Psychologie.
- Häcker, H., Schwenkmezger, P. & Utz, H. 1979. über die Verfälschbarkeit von Persönlichkeitsfragebogen und objektiven Persönlichkeitstests unter SD-Instruktionen und in einer Auslesesituation. *Diagnostica*, 25, 7-23.
- Hampel, R. & Klinkhammer, F. 1978. Verfälschungstendenzen beim Freiburger Persönlichkeitsinventar in einer Bewerbungssituation. *Psychologie und Praxis*, 22, 58-69.
- Hartley, J., Lindsey, D. & Burnhill, P. 1977. Alternatives in the typographic design of questionnaires. *Journal of Occupational Psychology*, 50, 299-304.
- Hartmann, H. 1972. *Empirische Sozialforschung*. München: Juventa.
- Hase, H. & Goldberg, R. 1967. Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin*, 67, 231-248.
- Hathaway, S. R. & Mc Kinley, J. C. 1963. *MMPI Saarbrücken, Handbuch*. Bern: Huber.
- Hayes, D. P. 1964. Item order on Guttman scales. *American Journal of Sociology*, 70, **51-58**.
- Heller, D. & Krüger, P. 1976. Analyse dreistufig zu beantwortender Fragebogenitems. *Psychologische Beiträge*, 18, 431-442.
- Hendrick, C., Borden, R., Giesen, M., Murray, E. J. & Seyfried, B. A. 1972. Effectiveness of ingratiation tactics in a cover letter on mail questionnaire response. *Psychonomic Science*, 26, 349-351.
- Hendrix, L. 1978. Studying ourselves: the questionnaire as a teaching tool. *Family Coordinator*, 27, 47-54.
- Hennig, W. 1971. Einige Fragen des Aufbaus von Interviewfragebogen und der Interviewer Ausbildung. In: Friedrich, W. (Hrsg.): *Methoden der marxistisch-leninistischen Sozialforschung*. Berlin: VEB Deutscher Verlag der Wissenschaften.
- Herrmann, T. & Stäcker, K. H. 1969. Sprachpsychologische Beiträge zur Sozialpsychologie. In: Graumann, C. F. (Hrsg.): *Handbuch der Psychologie*, Bd. 7: *Sozialpsychologie*, 1. Halbbd. Göttingen: Hogrefe.
- Hoeth, F. 1980. Antworttendenzen und ihre methodische Bedeutung für Befragungsv erfahren. In: Hartmann, K. D. & Koeppler, K. (Hrsg.): *Fortschritte der Marktpsychologie*, Bd. 2. Frankfurt: Fachbuchhandlung für Psychologie.
- Hoeth, F. & Köbler, V. 1967. Zusatzinstruktionen gegen Verfälschungstendenzen bei der Beantwortung von Persönlichkeitsfragebogen. *Diagnostica*, 13, 117-130.
- Holm, K. 1974a. Theorie der Frage. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 26, 91-114.

- Holm, K. 1974b. Theorie der Fragenbatterie. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 26, 316-341.
- Holm, K. (Hrsg.) 1975a. Die Befragung, Bd. 1. München: Francke.
- Holm, K. 1975b. Die Frage. In: Holm, K. (Hrsg.): Die Befragung, Bd. 1. München: Francke.
- Hornick, C. W., James, L. R. & Jones, A. P. 1977. Empirical item keying versus a rational approach to analyzing psychological climate questionnaire. *Applied Psychological Measurement*, 1, 489-500.
- Hornke, L. 1979. Konstruktion eines adaptiv-antwortabhängigen Fragebogens zur Erfassung der Prüfungsangst. *Diagnostica*, 25, 208-218.
- Janke, W. 1973. Das Dilemma von Persönlichkeitsfragebogen. Einleitung des Symposiums über Konstruktion von Fragebogen. In: Reinert, G. (Hrsg.): Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970. Göttingen: Hogrefe.
- Jannsen, J. P. 1978. Zur Validität und Reliabilität von Persönlichkeitsfragebogen in Ernstsituationen und beim Rollenspiel. Köln: TÜV Rheinland.
- Jetzschmann, H., Kallabis, H., Schulz, R. & Taubert, H. (Hrsg.) 1966. Einführung in die soziologische Forschung. Berlin: VEB Deutscher Verlag der Wissenschaften.
- Johnson, W. R., Sieveking, N. A. & Clanton, E. S. 1974. Effects of alternative positioning of open-ended questions in multiple-choice questionnaires. *Journal of Applied Psychology*, 59, 776-778.
- Jonsson, C. O. 1957. Questionnaires and interviews. Stockholm: Almqvist.
- Kahn, R. L. & Cannell, C. L. 1957. The dynamics of interviewing. New York: Wiley.
- Kalinowsky-Czech, M. 1979. Assoziationen und Entscheidungsprozesse bei der Beantwortung von Persönlichkeitsfragebogenitems. Bonn: Unveröff. Dipl. Arbeit.
- Kane, R. B. 1969. Computer generation of semantic-differential questionnaires. *Educational and Psychological Measurement*, 29, 191-192.
- Karmasin, F. & Karmasin, H. 1977. Einführung in die Methoden und Probleme der Umfrageforschung. Wien: Böhlau.
- Keil, W. 1973. Reaktionseinstellungen und Fragebogenkonstruktion. In: Reinert, G. (Hrsg.): Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970. Göttingen: Hogrefe.
- Kepes, S. Y. & True, J. E. 1967. Anonymity and attitudes toward work. *Psychological Reports*, 21, 353-356.
- Kinsey, A. C., Pomeroy, J. E. & Martin, C. E. 1970. Das sexuelle Verhalten des Mannes. Frankfurt: Fischer Bücherei.
- Kirschhofer-Bozenhardt, A. von & Kaplitza, G. 1975. Der Fragebogen. In: Holm, K. (Hrsg.): Die Befragung, Bd. 1. München: Francke.
- Kish, G. B. & Barnes, J. 1973. Variables that effect return rate of mailed questionnaires. *Journal of Clinical Psychology*, 29, 98-100.

- Knudsen, D. D., Pope, H. & Irish, D. P. 1967. Response differences to questions on sexual Standards: An interviewer - questionnaire comparison. *Public Opinion Quarterly*, 31, 290-297.
- König, R. (Hrsg.) 1972. *Das Interview*. Köln: Kiepenheuer & Witsch.
- König, R. (Hrsg.) 1974. *Handbuch der empirischen Sozialforschung*. Stuttgart: Enke.
- Koolwijk, J. von 1968. Fragebogenprofile. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 20, 780-791.
- Koolwijk, J. von 1969. 'Unangenehme Fragen'. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 21, 864-875.
- Koolwijk, J. von & Wieken-Mayser, M. (Hrsg.) 1974. *Techniken der empirischen Sozialforschung*, Bd. 4: Erhebungsmethoden: Die Befragung. München/Wien: Oldenbourg.
- Koomen, W. & Dijkstra, W. 1975. Effects of question length on verbal behavior in a bias-reduced interview Situation. *European Journal of Social Psychology*, 5, 399-403.
- Kraut, A. I., Wolfson, A. D. & Rothenberg, A. 1975. Some effects of position on opinion survey items. *Journal of Applied Psychology*, 60, 774-776.
- Kreutz, H. & Titscher, S. 1974. Die Konstruktion von Fragebögen. In: Koolwijk, J. von & Wieken-Mayser, M. (Hrsg.): *Techniken der empirischen Sozialforschung*, Bd. 4: Erhebungsmethoden: Die Befragung. München/Wien: Oldenbourg.
- Kuncel, R. B. 1973. Response processes and relative location of subject and item. *Educational and Psychological Measurement*, 34, 743-755.
- Lansing, J. B., Ginsberg, G. P. & Braaten, K. 1961. *An investigation of response error*. Urbana: University of Illinois.
- Lantermann, E. D. & Gehlen, H. 1977. Skalierung von Items und Individuen unter Beachtung individueller Urteilsstrukturen. *Zeitschrift für Sozialpsychologie*, 8, 242-246.
- Lazarsfeld, P. F. 1935. The art of asking ,why'. *National Marketing Review* 1. Zitiert nach: Maccoby, E. E. & Maccoby, N. 1972. *Das Interview: Ein Werkzeug der Sozialforschung*. In: König, R. (Hrsg.): *Das Interview*. Köln: Kiepenheuer & Witsch.
- Lazarsfeld, P. F. & Barton, A. H. 1955. Some general principles of questionnaire classification. In: Lazarsfeld, P. F. & Rosenberg, M. (eds.): *The language of social research*. Glencoe: The Free Press.
- Lennertz, E. 1973. Thesen zur Itemsammlung bei Persönlichkeitsfragebogen. In: Reinert, G. (Hrsg.): *Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970*. Göttingen: Hogrefe.
- Lienert, G. A. 1969. *Testaufbau und Testanalyse*. Weinheim: Beltz.
- Linsky, A. S. 1975. Stimulating responses to mailed questionnaires: A review. *Public Opinion Quarterly*, 39, 82-101.
- Litwak, E. 1956. A classification of biased questions. *American Journal of Sociology*, **62, 182-186.**

- Maccoby, E. E. & Maccoby, N. 1972. Das Interview: ein Werkzeug der Sozialforschung. In: König, R. (Hrsg.): Das Interview. Köln: Kiepenheuer & Witsch.
- Magnussen, D. 1966. Introduction to test theory. Reading: Addison-Wesley.
- Mauldin, W. P. & Marks, E. S. 1950. Problems of response in enumerative surveys. *American Sociological Review*, 15, 649-657.
- Mayntz, R., Holm, K. & Hübner, P. 1971. Einführung in die Methoden der empirischen Soziologie. Opladen: Westdeutscher Verlag.
- Mc Cormick, E. J., Jeanneret, P. R. & Mecham, R. C. 1969. The development and background of the position analysis questionnaire. Occupational Research Center Report No. 5. Lafayette: Purdue University Press.
- Mc Donagh, E. C. & Rosenblum, A. L. 1965. A comparison of mail questionnaires and subsequent structured interviews. *Public Opinion Quarterly*, 29, 131-136.
- Mc Kelvie, S. J. 1978. Graphic rating scales: how many categories? *British Journal of Psychology*, 69, 185-202.
- Metzner, H. & Mann, F. 1952. A limited comparison of two methods of data collection: The fixed alternative questionnaire and the open-ended interview. *American Sociological Review*, 17, 486-491.
- Metzner, H. & Mann, F. 1953. Effects of grouping related questions in questionnaires. *Public Opinion Quarterly*, 17, 136-141.
- Mittenecker, E. 1971. Subjektive Tests zur Messung der Persönlichkeit. In: Heiss, R., Groffmann, K. & Michel, L. (Hrsg.): Handbuch der Psychologie, Bd. 6: Psychologische Diagnostik. Göttingen: Hogrefe.
- Mucchielli, R. 1973. Die Befragung in der Sozialpsychologie. Salzburg: Müller.
- Münch, W. 1971. Datensammlung in den Sozialwissenschaften. Stuttgart: Kohlhammer.
- Narayana, C. L. 1977. Graphic positioning scale: an economical instrument for surveys. *Journal of Marketing*, 14, 118-122.
- Noelle, E. 1963. Umfragen in der Massengesellschaft. Reinbek: Rowohlt.
- Noelle-Neumann, E. 1970. Wanted: Rules for wording structured questionnaires. *Public Opinion Quarterly*, 34, 191-201.
- Noelle-Neumann, E. 1974. Probleme des Fragebogenaufbaus. In: Behrens, K. C. (Hrsg.): Handbuch der Marktforschung. Wiesbaden: Gabler.
- Nowakowska, M. 1971. A model for answering a questionnaire item. *Polish Psychological Bulletin*, 2, 37-45.
- Oppenheim, A. N. 1966. Questionnaire design and attitude measurement. New York: Basic Books.
- Osgood, C. E., Suci, G. J. & Tannenbaum, P. H. 1957. The measurement of meaning. Urbana: University of Illinois.
- Payne, S. L. 1951. The art of asking questions. Princeton: Princeton University Press.

- Perreault, W. D. 1975. Controlling order-effect bias. *Public Opinion Quarterly*, 39, 544-551.
- Phillips, B. S. 1966. *Social research: strategy and tactics*. New York: Macmillan.
- Phillips, B. S. 1970. *Empirische Sozialforschung*. Wien: Springer.
- Raab, E. 1974. Probleme der Frageformulierung. In: Behrens, K. C. (Hrsg.): *Handbuch der Marktforschung*. Wiesbaden: Gabler.
- Richardson, S. A., Dohrenwend, B. S. & Klein, D. 1965. *Interviewing: its forms and functions*. New York: Basic Books.
- Richter, H. J. 1967. Ist die schriftliche Befragung eine brauchbare Methode in der empirischen Sozialforschung? *Der Marktforscher*, 8, 234-235.
- Richter, H. J. 1969. *Grundlagen schriftlicher Massenbefragungen: ein verhaltens-theoretischer Beitrag zur Methodenkritik*. München: Phil. Diss.
- Ring, E. 1969. Haben Hintergrundfarben des Testmaterials Einfluß auf die Ergebnisse? *Psychologie und Praxis*, 13, 82-87.
- Ring, E. 1974. Wie man bei Listenfragen Einflüsse der Reihenfolge ausschalten kann. *Psychologie und Praxis*, 18, 105-113.
- Ring, E. 1975. Eine Fehlerquelle bei Bildern als Testvorlage. *Zeitschrift für experimentelle und angewandte Psychologie*, 22, 89-93.
- Rogers, T. B. 1974a. An analysis of the stages underlying the process of responding to personality items. *Acta Psychologica*, 38, 205-213.
- Rogers, T. B. 1974b. An analysis of two central stages underlying responding to personality items: the self-referent decision and response selection. *Journal of Research in Personality*, 8, 128-138.
- Rohrmann, B. 1978. Empirische Studien zur Entwicklung von Antwortskalen für die sozialwissenschaftliche Forschung. *Zeitschrift für Sozialpsychologie*, 9, 222-245.
- Rorer, L. G. 1965. The great response-style myth. *Psychological Bulletin*, 63, 129-156.
- Roslow, S., Wulfeck, H. & Corby, G. 1940. Consumer and opinion research: Experimental studies on the form of questions. *Journal of Applied Psychology*, 24, **334-346**.
- Rugg, D. 1941. Experiments in wording questions. *Public Opinion Quarterly*, 5, **91-92**.
- Rugg, D. & Cantril, H. 1972. Die Formulierung von Fragen. In: König, R. (Hrsg.): *Das Interview*. Köln: Kiepenheuer & Witsch.
- Scheier, I. H. & Cattell, R. B. 1965. Bestätigung von objektiven Testfaktoren und Beurteilung ihrer Beziehung zu Fragebogenfaktoren. *Diagnostica*, **11, 95-120**.
- Scheuch, E. K. 1962. Skalierungsverfahren in der Sozialforschung. In: König, R. (Hrsg.): *Handbuch der empirischen Sozialforschung*. Stuttgart: Enke.
- Scheuch, E. K. 1973. Das Interview in der Sozialforschung. In: König, R. (Hrsg.): *Handbuch der empirischen Sozialforschung*. Stuttgart: dtv.

- Scheuch, E. K. & Zehnpfennig, H. 1974. Skalierungsverfahren in der Sozialforschung. In: König, R. (Hrsg.): Handbuch der empirischen Sozialforschung. Stuttgart: Enke.
- Schneider, J. 1972. Versuchsleitereinfluß in Abhängigkeit von Merkmalen der Versuchsperson und dem Aussehen des Versuchsleiters. Saarbrücken: Phil. Diss.
- Schneider-Düker, M. & Schneider, J. 1977. Untersuchungen zum Beantwortungsprozeß bei psychodiagnostischen Fragebogen. Zeitschrift für experimentelle und angewandte Psychologie, 24, 282-302.
- Schreiber, K. 1974. Standardisierte und nichtstandardisierte Interviews. In: Behrens, K. C. (Hrsg.): Handbuch der Marktforschung. Wiesbaden: Gabler.
- Schriesheim, C. & Schriesheim, J. 1974. Development and empirical verification of new response categories to increase the validity of multiple response alternative questionnaires. Educational and Psychological Measurement, 34, 877-884.
- Schyberger, B. W. 1966. A case against direct questions on reading habits. Journal of Advertising Research, 6, 25-30.
- Sharma, S. N. & Singh, Y.P. 1967. Does the colour pull response? Manas: A Journal of Scientific Psychology, 14, 77-79.
- Sheatsley, P. B. 1972. Die Kunst des Interviewens. In: König, R. (Hrsg.): Das Interview. Köln: Kiepenheuer & Witsch.
- Sheth, J. & Roscoe, A. 1975. Impact of questionnaire length, follow-up methods, and geographical location on response rate to a mail survey. Journal of Applied Psychology, 60, 252-254.
- Sieber, M. 1979a. Zur Zuverlässigkeit von Eigenangaben bei einer Fragebogenuntersuchung. Zeitschrift für experimentelle und angewandte Psychologie, 26, 157-167.
- Sieber, M. 1979b. Zur Erhöhung der Rücksendequote bei einer postalischen Befragung. Zeitschrift für experimentelle und angewandte Psychologie, 26, 334-340.
- Simpson, R. H. 1944. The specific meaning of certain terms indicating different degrees of frequency. Quarterly Journal of Speech, 30, 328-330.
- Sixtl, F. 1972. Gedanken über die Verzahnung von allgemeiner und differentieller Psychologie. Archiv für Psychologie, 124, 145-157.
- Spoerer, E. 1979. Einführung in die Verkehrspsychologie. Darmstadt: Wissenschaftliche Buchgesellschaft.
- Steward, C. J. & Cash, W. B. 1978. Interviewing: principles and practics. Dubuque: W. C. Brown.
- Stollberger, R. 1966. Die Befragung. In: Jetzschmann, H., Kallabis, H., Schulz, R. & Taubert, H. (Hrsg.): Einführung in die soziologische Forschung. Berlin: VEB Deutscher Verlag der Wissenschaften.
- Strahan, R. & Gerbasi, K. C. 1973. Semantic style variance in personality questionnaires. Journal of Psychology, 85, 109-118.
- Strong, E. R. 1962. Good and poor interest items. Journal of Applied Psychology, 46, 269-275.

- Stroschein, F. R. 1965. Die Befragungstaktik in der Marktforschung. Wiesbaden: Gabler.
- Suchman, E. A. & Guttman, L. 1947. A Solution to the problem of question bias. *Public Opinion Quarterly*, 11, 445-455.
- Sudman, S. & Bradburn, N. M. 1974. Response effects in surveys: a review and Synthesis. Chicago: Aldine.
- Süllwold, F. 1969. Theorie und Methodik der Einstellungsmessung. In: Graumann, C. F. (Hrsg.): *Handbuch der Psychologie*, Bd. 7: Sozialpsychologie, 1. Halbbd. Göttingen: Hogrefe.
- Taetz, P. 1972. Conflicting group norms and the 'third' person in the interview. *American Journal of Sociology*, 68, 97-104.
- Terborg, J. R. & Peters, L. H. 1974. Some observations on wording of item stems for attitude questionnaires. *Psychological Reports*, 35, 463-466.
- Tholey, V. 1976. Die 'social desirability'-Variable bei der Beantwortung von Persönlichkeitsfragebogen. Darmstadt: Phil. Diss.
- Tittle, C. R. & Hill, R. J. 1967. The accuracy of self-reported data and prediction of political activity. *Public Opinion Quarterly*, 31, 103-106.
- Tränkle, U. 1974. Empirische Untersuchungen zum Einfluß der Befragungsmethode auf Befragungsergebnisse: standardisiertes Interview und schriftliche Befragung. Frankfurt: Unveröff. Jahresarbeit.
- Turner, C. & Fiske, D. W. 1968. Item quality and appropriateness of response processes. *Educational and Psychological Measurement*, 28, 297-315.
- Whitney, D. R. & Feldt, L. S. 1973. Analyzing questionnaire results: multiple tests of hypothesis and multivariate hypotheses. *Educational and Psychological Measurement*, 33, 365-380.
- Wieken, K. 1974. Die schriftliche Befragung. In: Koolwijk, J. von & Wieken-Mayser, M. (Hrsg.): *Techniken der empirischen Sozialforschung*, Bd. 4: Erhebungsmethoden: Die Befragung. München/Wien: Oldenbourg.
- Wildman, R. C. 1977. Effects of anonymity and social setting on survey responses. *Public Opinion Quarterly*, 41, 74-79.
- Wilk, G. 1974. Psychologische Probleme der Interviewsituation. In: Behrens, K. C. (Hrsg.): *Handbuch der Marktforschung*. Wiesbaden: Gabler.
- Willick, D. H. & Ashley, R. K. 1971. Survey question order and the political party preference of College students and their parents. *Public Opinion Quarterly*, 35, **189-199**.
- Wottawa, H. 1980. *Grundriß der Testtheorie*. München: Juventa.
- Wright, P. & Barnard, P. 1975. Just fill in this form: a review for designers. *Applied Ergonomics*, 6, 213-220.

6. Kapitel

Befragung

Ralf Schwarzer

1. Begriffsklärung und Übersicht

Die Befragung ist ein Spezialfall von Kommunikation, die in Abhebung vom Alltagsverständnis durch ihre wissenschaftliche Zielsetzung, den Grad der Strukturierung und Standardisierung sowie durch die damit verbundene Situationsdefinition charakterisiert ist, welche mit einer asymmetrischen Sozialbeziehung und einer einseitigen Verwertung der gewonnenen Information einhergeht. Das Methodenarsenal der Psychologie ist teilweise mit dem der empirischen Sozialforschung identisch, so daß insbesondere bei der Erörterung der Befragungsmethode auf die Erfahrungen mit der soziologisch orientierten Umfrageforschung zurückgegriffen werden kann. Innerhalb dieser Methode wird üblicherweise zwischen Interview und schriftlicher Befragung unterschieden. Scheuch (1973, 70) definiert: „Unter Interview als Forschungsinstrument sei hier verstanden ein planmäßiges Vorgehen mit wissenschaftlicher Zielsetzung, bei dem die Versuchsperson durch eine Reihe gezielter Fragen oder mitgeteilter Stimuli zu verbalen Informationen veranlaßt werden soll.“ Die schriftliche Befragung gilt bei ihm lediglich als Sonderform des Interviews (vgl. auch Atteslander 1969), während zum Beispiel bei Friedrichs (1973) beide Formen als eigenständige Methoden abgehandelt werden. Bei Verwendung der Befragung als Oberbegriff gelten folgende Merkmale. Es handelt sich meist um ein theoriegeleitetes, zumindest aber zielgerichtetes und regelhaftes Vorgehen der Datengewinnung; die Befragten werden entweder durch eine systematische Gesprächsoperation oder durch einen schriftlichen Fragenkatalog zu Informationsverarbeitungsprozessen veranlaßt, deren Resultat entweder verbal oder durch Antwortsymbole dem Forscher zur Verfügung gestellt wird. Definitionsgemäß sind damit andere Forschungsmethoden wie Experiment, Beobachtung und Inhaltsanalyse ausgeschlossen, was jedoch in der Forschungspraxis häufig durch Methodenvarianten und Methodenkombinationen wieder aufgehoben wird. Die begriffliche Abgrenzung der Befragung von anderen Methoden hat lediglich eine ordnungsstiftende und heuristische Funktion. Nicht eindeutig ist die Abgrenzung der Befragung von den Test- und Schätz-

verfahren, wozu auch die Persönlichkeitsfragebogen gezählt werden. Vom zu erfassenden Gegenstand her gesehen könnte man den Übergang von Meinungen zu dispositionalen Einstellungen als Nahtstelle der Verwendung der Befragung und des Persönlichkeitsfragebogens ansehen. Formal ließe sich diese Trennung durch nicht skalierte Auswertung bei der Befragung und skalierte Auswertung bei Tests oder testähnlichen Verfahren unterstützen. Damit wäre die Befragung im allgemeinen eine direkte Methode, die sich mit der Ebene der vorgefundenen beziehungsweise im Forschungsprozeß erzeugten Daten begnügt und auf die Schätzung latenter Merkmale verzichtet. Diese Auffassung läßt sich auch mit der gängigen Forschungspraxis begründen, wie sie in der auf Fakten und Meinungen gerichteten Umfrageforschung unter soziologischen Fragestellungen üblich ist. Für die Erfassung von überdauernden Einstellungen und anderen latenten Verhaltensdispositionen verfügt die Psychologie oft über bessere Methoden. Für die Erfassung von Kognitionen dagegen, die aufgrund ihrer Einmaligkeit, Prozeßhaftigkeit und Situationsspezifität mit anderen Instrumenten kaum zugänglich sind, erscheint die relativ anspruchsarme Befragungsmethode meist als vorteilhaft. Die psychologische Forschung nutzt hier vor allem das mündliche Interview als adaptive Gesprächsoperation, sowie methodische Varianten wie zum Beispiel das klinische Interview oder die Technik der kritischen Ereignisse (critical incidents technique). Grundsätzlich gilt, daß reine Formen der Befragung, wie sie in der empirischen Sozialforschung üblich sind, in der Psychologie seltener anzutreffen sind. Allerdings ist hier zu unterscheiden zwischen der psychologischen Forschung mit dem Ziel der Gewinnung generalisierbarer Erkenntnisse und der angewandten psychologischen Diagnostik mit dem Ziel indirekter Erfassung des Verhaltens und Erlebens. In der Diagnostik spielen Befragungsmethoden wie Anamnese und Exploration eine wichtige Rolle.

Wichtige Abgrenzungskriterien innerhalb der Vielzahl von Methodenvarianten sind die Strukturierung und die Standardisierung. Die Strukturierung betrifft den Aufbau der Befragung und läßt sich charakterisieren durch die Anordnung und den Grad der Geschlossenheit der Frage. Die Standardisierung betrifft die inhaltliche Vergleichbarkeit der Befragung und kommt in der gleichlautenden Frageformulierung und deren Reihenfolge für eine Mehrzahl von Befragten zum Ausdruck. Befragungen, die zugleich hoch standardisiert und hoch strukturiert sind, findet man zum Beispiel bei schriftlichen Massenumfragen (Surveys), bei denen die Fragebogen entweder über den Postversand oder durch Vorgabe in geschlossenen sozialen Einheiten (z.B. Schulen) in die Hände der Befragten gelangen. Eine niedrige Standardisierung findet man oft bei mündlichen Einzelinterviews, die entweder niedrig strukturiert (z.B. Exploration) oder hoch strukturiert (z.B. Erfassung bestimmter Denkprozesse) angelegt sein können. In manchen Fällen genügt ein Interviewleitfaden mit einigen Schlüsselfragen. Eine weitere Unterscheidung liegt in der informativischen, analytischen und diagnostischen Befragung (van Koolwijk 1974). Die

informativische Befragung ist auf die Erhebung von Fakten gerichtet, über die der Befragte als Informant bzw. Experte Auskunft geben kann. Die forensische Psychologie hat es bei der Zeugenaussage mit dieser Form zu tun (Arntzen 1970), die im allgemeinen wenig strukturiert und wenig standardisiert ist. Die analytische Befragung dient der Erfassung sozialer oder psychischer Gegenstände aus der Sicht der Befragten. Hoch standardisierte und hoch strukturierte Umfragen gehören dazu. Die diagnostische Befragung soll über die aktuellen und biographischen Aspekte der Persönlichkeit des Klienten Aufschluß geben, was mit Hilfe von Explorationen, Anamnesen und klinischen Interviews erreicht wird, die wenig standardisiert und wenig strukturiert sind.

Die instrumentellen Eigenschaften des Fragebogens oder Interviews hängen von der jeweiligen Konzeption der Befragung ab. Geht man informativisch vor, liefert der Wahrheitsgehalt der Information den Maßstab für die Validität. Geht man diagnostisch vor, wird die Angemessenheit der Aussagen im Hinblick auf tatsächliches Verhalten und Erleben der Klienten zu prüfen sein. Beim analytischen Vorgehen bietet es sich an, die Gütekriterien gemäß den Regeln der Testtheorie zu prüfen. Von besonderer Bedeutung ist die Überlegung, einen zu erfassenden Gegenstand mit Hilfe multipler Indikatoren zu untersuchen. Will man zum Beispiel erforschen, inwieweit jemand glaubt, mit seinen persönlichen Problemen fertig zu werden, ist es nicht sinnvoll, nur eine einzige Frage zu stellen oder auf eine einzige Aussage reagieren zu lassen. Bei einer Mehrzahl von Indikatoren besteht eher eine hinreichende Wahrscheinlichkeit dafür, das Konstrukt valide erfassen zu können. Fragt man dagegen nach der Parteizugehörigkeit oder Kinderzahl, reicht ein Indikator. Die Überprüfung der instrumentellen Eigenschaften bei Verwendung multipler Indikatoren läßt sich auf dem Wege über Strukturgleichungssysteme vornehmen (Jöreskog & Sörbom 1978), in denen jeder Indikator als beobachtete und jedes Konstrukt als latente Variable eingeht. Mit Hilfe der konfirmatorischen Faktorenanalyse läßt sich zum Beispiel das Meßmodell prüfen, indem zu jeder latenten Größe die theoretisch relevanten Indikatoren als kongenerische Variablen definiert werden. Wenn die Daten mit dem Modell verträglich sind, verfügt das Instrument über die erwünschten Eigenschaften. Der Ansatz läßt sich erweitern zu einem kombinierten Meß- und Kausalmodell wie zum Beispiel der Pfadanalyse mit latenten Variablen. Eine solche Vorgehensweise erscheint jedoch nur bei hoch standardisierten und hoch strukturierten Befragungen sinnvoll. Methodisch gibt es hier keinen Unterschied zu der Behandlung der instrumentellen Eigenschaften bei Persönlichkeitsfragebogen, die normalerweise dem Bereich der Test- und Schätzverfahren zugeordnet werden.

Zur Begriffsklärung gehört schließlich auch der zeitliche Aspekt bei der Anwendung der Befragungsmethode. In den meisten Fällen handelt es sich um Querschnittuntersuchungen, die innerhalb einer begrenzten Zeitspanne erfolgen. Bei einem längsschnittlichen Vorgehen ist zu unterscheiden zwischen der Panelstudie und der Trendstudie. Bei der Panelstudie werden dieselben Perso-

nen wiederholt befragt, während bei der Trendstudie zu jedem Meßzeitpunkt eine neue Stichprobe aus der ursprünglichen Population gezogen wird. In der Entwicklungspsychologie sind gegebenenfalls komplexere Versuchspläne von Vorteil, bei denen mehrere Panelstudien an verschiedenen Kohorten gleichzeitig durchgeführt werden (Baltes 1979).

2. Formen und Probleme der Befragung

2.1 Schriftliche Befragung

2.1.1 Vor- und Nachteile

Die mündlichen und schriftlichen Fragenvorgaben stellen die beiden Grundformen dar, wobei letztere zumeist als postalische Befragung verstanden wird (Bailey 1978). In der deutschen soziologischen Literatur wird dagegen die schriftliche Befragung meist nur als Sonderform des Interviews angesehen (Atteslander 1969, Scheuch 1973, Wicken 1974).

Die schriftliche Befragung verfügt über einige Vorteile. (1) *Kostenersparnis*. Da Herstellung und Versand von Fragebogen nur mit geringem Personalaufwand verbunden sind, ergibt sich im Vergleich zur personalintensiven - und manchmal reisekostenintensiven - mündlichen Befragung eine oft erhebliche Kostenersparnis, die auch bei besonders attraktiver Ausstattung des Materials günstig zu Buche schlagen kann. Es ist jedoch bei der Kalkulation zu berücksichtigen, daß bei extrem niedriger Rücklaufquote beträchtliche Porto- und Materialkosten ungenutzt bleiben. (2) *Zeitersparnis*. Wenn die Verfügbarkeit der Untersuchungsergebnisse eilbedürftig ist, ergeben sich mit dieser Befragungsform Vorteile, da alle Befragten die Unterlagen gleichzeitig erhalten können und man schon nach wenigen Tagen mit Beginn des Rücklaufs die Übertragung auf maschinenlesbare Datenträger beginnen kann. (3) *Bequemlichkeit für den Befragten*. Der Adressat kann sich Zeitpunkt und Umgebung für die Beantwortung des Fragebogens selbst auswählen. (4) *Anonymität*. Durch die Abwesenheit einer fragenden Person läßt sich das Vertrauen in die zugesicherte Anonymität erhöhen. (5) *Kein Interviewereinfluß*. Die systematischen Fehler, die beim mündlichen Interview als Elemente der sozialen Interaktionssituation auftreten, sind hier ausgeschaltet. (6) *Standardisierung*. Die schriftliche Vorgabe impliziert identische Frageformulierungen für alle Befragten, so daß durch gleiche Bedingungen die Durchführungsobjektivität gefördert wird, wenn auch konnotative Unterschiede insbesondere bei Begriffen mit semantischer Mehrdeutigkeit für Subpopulationen nicht auszuschließen sind. (7) *Informationssuche*. Der Befragte hat Zeit und Ruhe, um Fragen länger zu durchdenken und sich der Richtigkeit seiner Antworten zu vergewissern. So kann er bei Fragen, die sich zum Beispiel auf Fakten in seiner Biographie beziehen,

Angehörige konsultieren. Allerdings werden hier durch manche Fragen länger dauernde Reflexionsprozesse stimuliert, die zu Antworten führen, welche ohne das Forschungsinstrument nicht zustande gekommen wären. Diese Reaktivität kann entweder als unerwünschtes Untersuchungsartefakt oder als erwünschte Subjektveränderung in Handlungsforschungsprojekten aufgefaßt werden. (8) *Zugänglichkeit*. Regional verstreut lebende Adressaten werden vor allem unter dem Zeit- und Kostengesichtspunkt am besten durch postalische Befragung zugänglich.

Andererseits sind schwerwiegende Nachteile der schriftlichen Befragung zu verzeichnen. (1) *Rücklaufquote*. Die Zahl der ausgefüllt zurückgesandten Fragebogen ist manchmal sehr gering, und ein Anteil von 50% wird schon als angemessen bezeichnet (Babbie 1973). Damit kann zugleich eine Stichprobenverzerrung gegeben sein. Versuche, solche Verzerrungen aufzudecken, richten sich meist auf die Prüfung der Verteilung einfacher Merkmale (z.B. Geschlecht) in der Reststichprobe. Die theoretisch wichtige Frage nach einer Stichprobenverzerrung im Hinblick auf den Untersuchungsgegenstand selbst dürfte jedoch nicht zu beantworten sein. Weiterhin bleibt ungeklärt, welche Ausfälle auf Adressenfehler oder auf absichtliche Nichtbeantwortung zurückzuführen sind. (2) *Unvollständigkeit*. Mangels der Kontrolle eines anwesenden Interviewers werden viele Fragen nicht beantwortet. Statt dessen bringt der Befragte seinen ungedämpften Zorn über mißverständene oder provokante Fragen durch an den Rand geschriebene Schimpfwörter zum Ausdruck. Bei der statistischen Auswertung kann es somit zu einem verheerenden Datenschwund kommen, falls es erforderlich ist, Prozeduren mit fallweisem Ausschluß zu verwenden. (3) *Unkontrollierbarkeit der Erhebungssituation*. Man kann nicht sicher sein, ob wirklich die angeschriebene Person oder jemand anders den Fragebogen ausfüllt. Außerdem können situative Merkmale wie Ablenkung durch Lärm oder Kontaktpersonen die Validität einschränken. Damit verbunden ist die fehlende Kontrolle über die Wahl der Reihenfolge und den Zeitpunkt der Bearbeitung der Fragen. (4) *Unsichtbarkeit*. Es erfolgt eine Reduktion auf Verbalverhalten beziehungsweise abstrakt-symbolische Handlungsergebnisse (Ankreuzen), während nonverbales Verhalten, welches in der mündlichen Interviewsituation Validitätshinweise geben kann, aufgrund der Unsichtbarkeit des Befragten ausgeschlossen wird (Scherer 1974). Auch Spontanäußerungen bleiben unerfaßt. (5) *Keine Flexibilität*. Es ist kein Interviewer anwesend, der durch Nachfragen Informationen gewinnen, Fragen erläutern, Motivation fördern oder zornig erregte Personen beruhigen kann.

2.1.2 Weitere Probleme und Besonderheiten

Die Vor- und Nachteile können in Abhängigkeit des jeweiligen Untersuchungsgegenstandes und Handlungskontextes als unterschiedlich gravierend

angesehen werden. Es gibt eine Fülle von Arbeiten, die sich mit den genannten Aspekten und anderen Einflußfaktoren auf Ergebnisverzerrungen von schriftlichen Befragungen befassen. Die meisten richten sich auf Experimente zur Bestimmung von Variablen, die die Rücklaufquote beeinflussen. Die Antwortbereitschaft erscheint demnach als abhängig von der vermeintlichen Seriosität des Absenders (z.B. von der Regierung geförderte Forschung) und dessen Begleitschreiben (Scott 1961), von der Länge des Fragebogens (Berdie 1973), der Farbe des Fragebogens (Matteson 1975), von finanziellen Anreizen (Armstrong 1975), von handschriftlichen Zusätzen und anderen Mitteln der Personalisierung (Carpenter 1975), von der Wahl der Briefmarken (Hensley 1974), der zugesicherten Anonymität (Fuller 1974) und weiteren Faktoren (Linsky 1975, Sudman/Bradburn 1974). Außerdem wird untersucht, ob es Persönlichkeitsunterschiede zwischen Antwortern und Nichtantwortern gibt. Binder, Sieber und Angst (1979) fanden heraus, daß Personen mit hohem Berufs- und Bildungsstatus sowie Nichtraucher häufiger den Fragebogen zurücksandten. Die Antworter waren hinsichtlich der Skalen des Freiburger Persönlichkeitsinventars gehemmter, weniger gesellig und weniger dominant, aber zugleich offener als die Nichtantworter. Die Effekte waren sehr schwach. Die Erforschung von Bedingungen für die Rücklaufquote ist bisher überwiegend theorielos und methodisch relativ anspruchslos vorgenommen worden. Es fehlen genügende mehrfaktorielle und multivariate Untersuchungen, die auch Auskunft über Moderator- bzw. Interaktionseffekte geben und die in theoretischer Hinsicht auf sozialpsychologische Erklärungen (z.B. Altruismus) zurückgreifen können. Nach einer anspruchsvollen Reanalyse von 98 Experimenten zur Rücklaufquote kommen Heberlein und Baumgartner (1978) zu der Auffassung, daß nicht die niedrige Rücklaufquote selbst das Problem darstellt, sondern ihre Variabilität im Hinblick auf verschiedene Forscher, Populationen, Fragebogen und Vorgehensweisen. Als wichtigster Faktor wird die subjektive Bedeutsamkeit angesehen, die durch eine Vielzahl von Anreizen, wie sie schon immer in der Literatur diskutiert worden sind, seitens des Forschers erhöht werden kann. Unter dieser Perspektive erscheinen die widersprüchlichen Befunde bezüglich der Länge des Fragebogens in einem anderen Licht. Offenbar wird bei einem langen Fragebogen der durch erhöhten Arbeitsaufwand gegebene negative Effekt mehr als ausgeglichen durch die Unterstellung von Bedeutsamkeit, die ihm zugeschrieben wird. Sehr kurze Fragebogen werden möglicherweise als weniger bedeutsam erlebt.

In der Psychologie ist die postalische Variante der schriftlichen Befragung weniger gebräuchlich als zum Beispiel in der Markt- und Meinungsforschung. Das mag auch damit zusammenhängen, daß Auskünfte über Tatsachen psychologisch weniger interessant erscheinen als die Erfassung subjektiver Wahrnehmungen eigenen und fremden Erlebens und Handelns. Beide Aspekte verdienen neuerdings mehr Beachtung im Rahmen der Erfassung von Auswirkungen kritischer Lebensereignisse. Mit dem Life Experiences Survey (LES) zum

Beispiel wird danach gefragt, ob innerhalb der letzten sechs oder zwölf Monate ein Verlust von Angehörigen, eine Verschuldung, eine Trennung, eine Arbeitslosigkeit, eine Krankheit usw. eingetreten sind und als wie beeinträchtigend dieses Ereignis erlebt worden ist, was auf einer siebenstufigen Skala anzugeben ist (Johnson/Sarason 1979). Solche Befragungen dienen der quantitativen Abschätzung von subjektiven Fehlanpassungen und situativ bedingten Streßreaktionen.

Die schriftliche Befragung erfolgt in der psychologischen Forschung meistens als Gruppenbefragung bzw. als „Befragung unter Aufsicht“ (Anger 1969), was in der Regel zu einer Rücklaufquote von 100% führt. Man bedient sich dabei der leichten Zugänglichkeit von Schulen, Betrieben, militärischen Einheiten, Krankenhäusern usw., was jedoch normalerweise mit dem Verzicht auf Zufallsstichproben verbunden ist und für die Wahl der Analyseeinheit (Individuen oder Aggregate) Probleme aufwirft. Sofern es aus theoretischer Sicht vernünftig ist, werden daher Mehrebenenanalysen empfohlen. Die Ausnutzung institutioneller Gegebenheiten für Befragungszwecke beschränkt sich nicht nur auf die dort erfaßte Population, sondern auch auf Angehörige. So werden in der Pädagogischen Psychologie zum Beispiel nicht nur Schüler, sondern auch deren Eltern und Geschwister befragt, indem man die Lehrer als Verteiler und die Schüler als Boten der Fragebogen einsetzt, so daß der Rücklauf trotz gewahrter Anonymität exakt kontrolliert werden kann. Im Zusammenhang mit der Schulversuchsbegleitforschung oder Curriculumevaluation lassen sich auf diese Weise ganze soziale Einheiten und Umfeldler mit nahezu kompletten Daten erfassen (Schwarzer 1975).

2.2 Die mündliche Befragung

2.2.1 Vor- und Nachteile

Die mündliche Befragung wird im allgemeinen als Interview bezeichnet. Gegenüber der schriftlichen Befragung verfügt sie über einige Vorteile. (1) *Flexibilität*. Der Interviewer kann sich den Bedürfnissen des Befragten anpassen, indem er Mißverständnisse ausräumt und das mit der Frage Gemeinte noch einmal umgangssprachlich erläutert. (2) *Spontaneität*. Die impulsiven Reaktionen des Befragten, die manchmal valider sind als die wohlüberlegten Reaktionen, können vom Interviewer festgehalten werden. Der geschulte Interviewer kann aufgrund des mit der Beantwortung verbundenen Gesamteindrucks darüber entscheiden, welcher Kategorie eine Reaktion zuzuordnen ist. (3) *Nicht-verbale Reaktionen*. Der Interviewer kann auch die nonverbalen und paralingualen Verhaltensweisen beobachten wie Achselzucken, Lachen, Erröten usw. und damit den Grad der Validität der Aussagen abschätzen. (4) *Identifikation*. Der Befragte ist eindeutig als solcher identifiziert. Er kann die Beantwortung

nicht an andere Personen delegieren, wie es bei der postalischen Befragung möglich ist. (5) *Kontrolle der Erhebungssituation*. Der Interviewer kann für alle Befragten eine vergleichbare entspannte Atmosphäre schaffen und dafür sorgen, daß die Fragen in Ruhe und ohne unerwünschte Anwesenheit Dritter beantwortet werden. (6) *Reihenfolge*. Es wird sichergestellt, daß die vorgegebene Reihenfolge der Fragen eingehalten wird und der Befragte nicht beliebig von einer Frage zur anderen springen kann, wodurch sich Positionseffekte ergeben könnten. (7) *Komplexität*. Bei komplexen Fragenkatalogen mit vielen Filtern ist das mündliche Interview überlegen, weil der trainierte Interviewer die vielen Sprungbefehle im Kopf hat und den Befragten sicher durch das Gestrüpp führen kann. (8) *Dauer*. Die Bearbeitungsdauer läßt sich registrieren und als zusätzliche Variable in die Auswertung aufnehmen. Dies kann ein Indikator für den Grad der persönlichen Involviertheit mit dem Thema sein. (9) *Vollständigkeit*. Der Interviewer vergewissert sich der vollständigen Bearbeitung aller Fragen und wird bei verweigerter Beantwortung einer Frage prüfen, ob eine echte „Meinungslosigkeit“ oder nur Bequemlichkeit vorliegt. (10) *Rücklauf*. Die Stichprobe bleibt meistens zu mehr als 80% erhalten, da die Zahl der Verweigerer und die der nicht Auffindbaren gering ist, was allerdings auch von einigen Faktoren abhängig ist, die zu einer Variation der Antwortbereitschaft führen (z.B. heikles Thema). Analphabeten sind für Interviews zugänglich. Viele Personen sind eher bereit, eine mündliche als eine schriftliche Interaktion einzugehen.

Den Vorteilen stehen einige Nachteile gegenüber. (1) *Kostenaufwand*. Die Personalkosten für den Interviewerstab während der Trainingsphase und der Feldarbeit schlagen zu Buche. Manchmal treten Reisekosten hinzu. Aufgrund regionaler Verstreutheit ist unter diesen Umständen manchmal kein Zugang zu der interessierenden Personengruppe möglich. (2) *Zeitaufwand*. Hat man große Stichproben und wenige Interviewer, muß die Befragung auf einen längeren Zeitraum verteilt werden. Da die zu Befragenden nicht immer erreichbar oder antwortunwillig sind, gibt es zusätzliche Verzögerungen. Die daraus resultierende späte Verfügbarkeit über die Forschungsergebnisse ist nur ein Nachteil. Noch gravierender dürften eventuelle Meßzeitpunkteffekte sein, da während der Befragungsmonate individuelle oder kollektive Ereignisse eintreten können, die die Beantwortungsrichtung beeinflussen. (3) *Geringere Anonymität*. Das persönliche Gegenüber von Interviewer und Befragten reduziert die zugesicherte Anonymität und kann als Bedrohung empfunden werden, die zu einer Verfälschung der Antworten oder zur Teilnahmeverweigerung führt. (4) *Belästigung*. Wird der Befragte zu Hause oder am Arbeitsplatz aufgesucht, kann er dies als belästigend empfinden, was die Befragungssituation beeinträchtigt, die dadurch zu einer Streßsituation werden kann. Vor allem wenn zusätzlich Dritte anwesend sind, die die Interaktion beobachten, können soziale Ängstlichkeit, soziale Erwünschtheit oder Imponiergehabe den Dialog beeinflussen. (5) *Interviewereinfluß*. Persönliche Merkmale des Interviewers wie Geschlecht,

äußere Erscheinung, Auftretensweise, Alter usw. können systematische Fehler (bias) in den Rapport einschleusen. (6) *Geringe Standardisierung*. Der Vorteil der Flexibilität des Interviewers stellt zugleich ein Problem dar, weil das individuelle Erläutern und nondirektive Nachfragen die Vergleichbarkeit zwischen den Interviews beeinträchtigt.

2.2.2 *Der Interviewer*

Das Interview stellt eine besondere soziale Interaktion dar, in der ein Interaktionspartner auftragsgemäß fragt und der andere freiwillig antwortet. Die Kommunikationssituation ist künstlich, asymmetrisch und von sehr kurzer Dauer. Der Interviewer verfolgt im Dialog keine persönlichen Interessen, sondern die seines Auftraggebers. Er übernimmt eine spezifische Rolle im Forschungsprozeß, die zwischen dem Forscher und dem Befragten agiert. Daher erhält der an den Fragenkatalog und zusätzliche Instruktionen gebundene Interviewer methodisch gesehen den Status eines Instruments. Den Menschen als Forschungsinstrument einzusetzen, ist eine riskante Angelegenheit. Einerseits verfügt er über eine Informationsverarbeitungskapazität ohnegleichen, die ihm Flexibilität im Verfolgen der Ziele erlaubt, andererseits stellt er eine Quelle von systematischen Fehlern dar. Der zweite Aspekt ist seit Jahrzehnten Gegenstand von soziologisch orientierten Untersuchungen (Bailey 1978, Erbslöh/Wiendieck 1974). Als verantwortlich für diesen „interviewer bias“ lassen sich Merkmale der äußeren Erscheinung, latente Verhaltensdispositionen und situationsspezifische Verhaltensweisen unterscheiden. Die älteren amerikanischen Studien richten sich vor allem auf den Einfluß von Rassen-, Sozialschicht- und Geschlechtszugehörigkeit, Alter und Kleidung. Danach erzeugen Interviewer mit unterschiedlicher Kategorienzugehörigkeit unterschiedliche Effekte bei Personen, die entweder derselben oder der entgegengesetzten Kategorie angehören. Die Auswirkungen sind an der Zahl, der Länge und der Ehrlichkeit der Antworten ablesbar. Psychologisch gesehen ist diese Forschungstradition unbefriedigend, weil äußere Merkmale keine direkte Kausalwirkung auf das Verhalten des Dialogpartners ausüben. Vielmehr wäre es von Bedeutung, die kognitiven Zwischenprozesse zu analysieren, welche für Verhaltensänderungen verantwortlich sind. Diese Kritik gilt in abgeschwächter Weise gleichermaßen für die Untersuchungen der Einflüsse aufgrund von Persönlichkeitseigenschaften (latenten Verhaltensdispositionen) von Interviewern und aufgrund von spezifischen Verhaltensweisen des Interviewers während der Befragung. In der sozialen Interaktion sind Kognitionen verhaltenswirksam (Frey 1978), und andere Merkmale dienen lediglich als Indikatoren oder als zusammenfassende Konstrukte. Die Analyse von Interviewereffekten bedarf demnach der besonderen Berücksichtigung von sozialen Vergleichsprozessen (Suls/Miller 1977). Der Befragte vergleicht sich selber mit dem Interviewer und definiert auf diesem Wege die wahrgenommene soziale Distanz zwi-

schen beiden Personen. Dabei verwendet er äußere Merkmale, die vermutete Sozialschichtzugehörigkeit und die angenommene Einstellungskongruenz als Hinweisreize. Die Definition der Distanz bzw. des Ähnlichkeitsgrades bezieht sich offenbar vor allem auf solche Merkmale, die mit der Thematik des Interviews verwandt sind (related-attribute similarity). Danach wäre es zum Beispiel in einem Interview über Einstellungen zu Gastarbeitern von Bedeutung, ob der Interviewer z.B. wie ein Student aus der Dritten Welt oder wie ein deutscher Landwirt aussieht. Dagegen wird der entsprechende Interviewereffekt bei den Themen Schulreform oder Flugsicherheit geringer sein. Das Ähnlichkeitskonzept hängt mit der subjektiv wahrgenommenen sozialen Dominanz zusammen. Die Unähnlichkeit zwischen Interviewer und Befragten kann in Abhängigkeit vom Befragungsgegenstand oder einem weitergehenden Befragungskontext relevant oder irrelevant sein, entscheidend ist jedoch, ob diese Unähnlichkeit zugleich ein soziales Dominanzgefälle impliziert. Die Wahrnehmung eines dominanten Interviewpartners kann zu einer kognizierten Bedrohung führen, welche beim Befragten die Auswahl solcher Antworten begünstigt, die geeignet erscheinen, den Grad der sozialen Bedrohlichkeit zu reduzieren. Dazu gehören Antworten defensiver oder aggressiver Art bzw. sozial erwünschte Äußerungen, da Konformität streßreduzierend wirken kann. Die perzipierte Unähnlichkeit im Sinne von sozialer Dominanz muß allerdings nicht in jedem Falle nur auf die Person des Interviewers zurückzuführen sein. Der hinter dem Interviewer stehende Auftraggeber kann durch diesen hindurch wirken (sponsorship bias) und effektvermindernd oder effektvergrößernd sein. Tritt ein konservativ erscheinender Interviewer zum Beispiel im Auftrag einer Gewerkschaft auf, so könnte dies eine Verminderung des Interviewereffekts bedeuten. Je nach Forschungsgegenstand und -interesse können Interviewereffekte im Sinne der Fragestellung genutzt werden. Für eine „harte“ Befragung eignen sich dominante Interviewer, während für eine „weiche“ Befragung Interviewer mit großer Ähnlichkeit zum Befragten oder sogar tendenzieller Submissivität gesucht werden müssen, um die Voraussetzungen für ein nondirektives Vorgehen zu schaffen. Im Regelfall wird man jedoch eine neutrale Befragung anstreben.

Das Ziel der Interviewerschulung sollte daher darin liegen, möglichst neutrales Auftreten zu erzeugen und interindividuelle Unterschiede zwischen den Interviewern zu reduzieren. Es erscheint sinnvoll, sie mit einer rollenadäquaten „Uniform“ zu versehen, um sie soweit wie möglich zu entpersonalisieren. Im Idealfall treten sie dem Befragten nicht als Persönlichkeit, sondern als Forschungsinstrument entgegen.

2.2.3 Der Befragte

Die Trennung von Effekten auf der Seite des Interviewers von denen auf der Seite des Befragten ist nicht so sehr systematisch als vielmehr heuristisch. Es ist

im Einzelfall normalerweise nicht zu entscheiden, ob die Datenverzerrung auf eine Person alleine zurückzuführen ist, da es sich um einen sozialen Interaktionsprozeß handelt, in dem die Daten erzeugt werden. An den Befragten wird eine Menge von Erwartungen herangetragen wie zum Beispiel die kurzfristige Bereitschaft, Antworten zu geben, ohne selbst zu fragen, und Informationen über sich und andere preiszugeben, ohne daß dieses Verhalten unmittelbare Konsequenzen innerhalb der alltäglichen Lebenswelt nach sich zieht, wie es bei anderen sozialen Interaktionen der Fall ist. Der Befragte wird als Datenträger oder potentieller Datenproduzent angesehen, wobei davon ausgegangen wird, daß er diese Attribution für die eigene Rollendefinition übernimmt. Damit wird dem Interviewer weitgehend die Kontrolle der Situation überlassen. Metakommunikation gilt während der Dauer des Interviews als unzulässig oder störend. Daraus ergeben sich vor allem dann Probleme, wenn die Befragten sich freiwillig auf die Situation einlassen, weil damit in der Regel die Stichprobe nicht für die angestrebte Population repräsentativ ist. Eine „Psychologie des Freiwilligen“ würde ermitteln, daß seine Altruismustendenz überdurchschnittlich hoch liegt oder er seine eigene Sprach- und Sozialkompetenz hoch einschätzt und unter Beweis stellen möchte usw. Wie sich die Gruppe der Befragten rekrutiert, erscheint danach nicht nur als ein quantitativ kalkulierbares Problem des Auswahlverfahrens, sondern mindestens ebenso als ein motivationales Problem. Dabei kann es eine Rolle spielen, ob der Befragte eine angenehme Sozialbeziehung zum Interviewer entwickeln beziehungsweise aufrechterhalten möchte, ob er zum Hilfehandeln motiviert ist oder ob er aufgrund intellektuell gewandten Handelns seine Selbsteinschätzung überprüfen oder erhöhen möchte und die Interviewsituation als günstige Gelegenheit dafür ansieht. Als typische Antwortverzerrungen (response sets) werden dann soziale Erwünschtheit und Zustimmungstendenz (Akquieszenz) hervorgebracht. Die Probleme sind hier nicht viel anders als jene, die in den letzten Jahrzehnten ausführlich im Zusammenhang mit Persönlichkeitsfragebogen diskutiert worden sind (Esser 1977, Janke 1973).

Das Auftreten von Antworttendenzen ist verwandt mit der Tendenz zur Nichtbeantwortung. Es wird unterschieden zwischen Frageverweigerung, Nichtinformiertheit, Meinungslosigkeit und Unentschiedenheit (Esser 1974), ohne daß hier immer eine genaue Trennung möglich wäre. Die Nichtinformiertheit läßt sich durch eine Filterfrage kontrollieren und dadurch abgrenzen von der Meinungslosigkeit. So haben zum Beispiel Schumann und Presser (1978) den Einfluß von Filtern auf die Nichtbeantwortung von Meinungsfragen untersucht. Dagegen ist die echte Unentschiedenheit des Befragten über die Verwendung von Filtern oder Neutralkategorien nicht vollkommen erfäßbar, weil kognitive Prozesse eine Rolle spielen, die möglicherweise nicht auf den Inhalt der Frage, sondern auf die Bedrohlichkeit einer Alternativentscheidung beziehungsweise auf die Interaktionssituation insgesamt gerichtet sein können, so daß die Selbstbezogenheit der Reizkonfiguration zur Fehlerquelle

wird (self-serving bias). Derselbe Mechanismus führt auch dazu, daß gelegentlich Einstellungen geäußert werden, die sachlich nicht möglich sind. Werden zum Beispiel Studenten nach ihrer persönlichen Meinung zum „Psychologengesetz“ gefragt, welches der Bundestag angeblich im letzten Jahr verabschiedet hat, so findet man tatsächlich einige, die bereitwillig und sogar ausführlich ihre Meinung dazu artikulieren. Es geht in solchen Situationen nicht um den Inhalt, sondern eher um ein rollenadäquates intellektuelles Verhalten. Antwort- und Nichtantworttendenzen entstehen offenbar im Zusammenhang mit Kognitionen, die die eigene Person betreffen. Es handelt sich bei den Daten daher um eine Konfundierung von zwei Reaktionsweisen. Der Befragte ist sowohl reaktiv gegenüber dem Inhalt als auch gegenüber der Erhebungssituation insgesamt, wobei insbesondere der Interviewer als einer von vielen situativen Stimuli wirksam ist. Die doppelte Reaktivität führt zu einer Einschränkung der Validität der Messung. Gelegentlich wird daher vorgeschlagen, nichtreaktive Verfahren (z.B. Beobachtung und Inhaltsanalyse) anstelle der Befragung zu verwenden. Das wäre jedoch eine unnötige Reduzierung einer prinzipiell multi-methodischen Forschungsstrategie. Vielmehr geht es darum, die Validitätsprobleme transparent zu machen und in jedem Einzelfall die differentielle Gültigkeit der Daten zu maximieren, das heißt nach Beachtung der üblichen Regeln zu ermitteln, für welche Situationen und bei welchen Personen die Validität höher oder geringer ausfällt und welche psychischen Prozesse aus theoretischen Gründen dafür verantwortlich gemacht werden können. Dieser Aspekt wird bei der herkömmlichen Umfrageforschung - aus verständlichen Gründen - im allgemeinen vernachlässigt.

2.3 Einige Sonderformen

2.3.1 Realkontakt-Befragung

Angesichts der Validitätsprobleme sind Vorschläge gemacht worden, um die Erhebungssituation grundsätzlich zu verändern. So stammt von Kreutz (1972) der Vorschlag, zusätzlich zu den üblichen Forschungskontakt-Befragungen vor allem sogenannte Realkontakt-Befragungen durchzuführen. Dabei übernimmt der Interviewer eine Rolle, die in dem Untersuchungsfeld natürlicherweise bereits vorhanden ist. Wenn es zum Beispiel darum geht, das Interaktionsverhalten oder die Leistungsanforderungen von Hochschullehrern zu untersuchen, kann man Studenten in deren Sprechstunden schicken und Fragen stellen lassen, deren Beantwortung ohne Verzerrung zu den entsprechenden Daten führt. Das Anwendungsfeld der Realkontakt-Befragung ist umfangreich (Kreutz 1972, S. 111). Die Interviewer können als Käufer auftreten, um das Verkaufsverhalten in verschiedenen Branchen zu erforschen; sie können als Wohnungssuchende auftreten, um das Verhalten von Besitzern und Maklern zu studieren; sie können als Stellenbewerber auftreten, um Selektionsvorgänge

und Anforderungskriterien zu ermitteln; sie können als Kranke auftreten, um das Verhalten von Ärzten zu registrieren usw. Das Vorgehen läßt sich verbinden mit anderen Sonderformen wie zum Beispiel dem Tandem-Interview, bei dem jemand von zwei Interviewern zugleich befragt wird. Will man die Interaktionsvorgänge in einer psychologischen Eheberatungsstelle untersuchen, bietet es sich an, die Interviewer als Ehepaare zu tarnen, die mit einem erforderlichen Leidensdruck die Beratungsstelle aufsuchen. Die Realkontakt-Befragung nähert sich anderen Methoden wie zum Beispiel der verdeckt-teilnehmenden Beobachtung. Bei isolierender Bedingungsvariation ist der Übergang zu einer experimentellen Anordnung gegeben. Der entscheidende Vorteil des Verfahrens liegt in dem Versuch, Validität zu erhöhen, indem die erwünschte Reaktivität auf den Inhalt maximiert und die unerwünschte Reaktivität auf das Meßinstrument minimiert wird. Da der Interviewer seinen tatsächlichen Auftrag und seine Rolle im Forschungsprozeß nicht zu erkennen gibt, handelt es sich hier um eine verdeckte Methode, deren forschungsethische Implikationen mit zu reflektieren sind.

2.3.2 *Telefoninterview*

Das Telefoninterview dient dazu, in einer aktuellen Situation billig und schnellstmöglich ein vorläufiges Meinungsbild zusammenzutragen (Bailey 1978). Es ist daher für Meinungsforschungsinstitute und Tageszeitungen gelegentlich brauchbar, während es in der psychologischen Forschung bisher keine bedeutende Rolle spielt. Andererseits sind die potentiellen Vorteile dieser Variante noch nicht genügend ausgeschöpft worden. Innerhalb weniger Stunden nach Eintreten von zum Beispiel Unglücken oder politischen Ereignissen ist bereits eine kostengünstige Datenerhebung möglich, ohne daß zwischenzeitlich ein öffentlicher Meinungsbildungsprozeß durch Massenmedieneinflüsse und Gruppendiskussionen wirksam geworden ist. Ein entscheidender Nachteil liegt in der Reduzierung der Stichprobe auf Telefonbesitzer und solche, die gerade in der Wohnung anwesend sind. Weiterhin ist mit einer großen Verweigerungsquote zu rechnen, da Telefoninterviews nicht alltäglich sind und als Telefonterror mißverstanden werden könnten. Die Seriosität des Unternehmens ist wegen fehlender Legitimationsmöglichkeiten fragwürdig. Diese Nachteile können überwunden werden, wenn von vornherein nur eine Population untersucht werden soll, die telefonisch erreichbar ist, und wenn das Telefoninterview vorher schriftlich vereinbart worden ist. Bei zum Beispiel einer Evaluation von Problemen, die in psychologischen Schulforschungsprojekten auftreten, kann der Evaluator in einem Schreiben auf Kopfbogen die verschiedenen Projektleiter um ein Telefoninterview bitten, für das ein Zeitpunkt vorgeschlagen wird und für das die wichtigsten Fragen schriftlich vorgegeben sind („Werden Sie in Zukunft Rattenexperimente durchführen, wenn der Zugang zu Schulen nicht mehr möglich sein sollte, oder was sonst?“). Eine

andere wenig genutzte Möglichkeit liegt in der Verwendung des Telefoninterviews bei Panelstudien. Man untersucht zum Beispiel im Längsschnitt die Stabilität und Veränderung von Einstellungen gegenüber Energie- und Umweltproblemen. Nach einem persönlichen Erstinterview wird die gewonnene Stichprobe später nur noch telefonisch nachbefragt, wobei die Untersuchung den Charakter eines Zeitreihenexperiments erhalten kann, wenn man die natürlich auftretenden Ereignisse als Treatments betrachtet. Sobald zum Beispiel eine Umweltkatastrophe in den Massenmedien gemeldet wird, greift der Forscher zum Telefon und realisiert einen neuen Meßzeitpunkt in seiner Panelstudie.

2.3.3 Kinderinterview

In der Entwicklungspsychologie richten sich viele Forschungsfragen auf die Entstehungsbedingungen von Dispositionen, Werten und Einstellungen während der familiären und schulischen Sozialisation. Der Mangel an ökonomisch einsetzbaren Skalen für Kinder führt zur Anwendung weniger standardisierter und strukturierter Verfahren. Kinderinterviews müssen die mangelnde Sprachbeherrschung, das fehlende Abstraktionsniveau, die kurze Aufmerksamkeitsspanne, die Ideenflüchtigkeit, die Unvertrautheit mit der Erhebungssituation und die Besonderheiten der Kind-Erwachsener-Beziehung berücksichtigen (Bailey 1978). Bei Vorschulkindern und Erstkläßlern sind schriftliche Untersuchungsverfahren ausgeschlossen. Man behilft sich mit dem Vorlesen von Fragen und mit Veranschaulichungstechniken und fordert vom Kind sehr einfache Antworten, die manchmal auch nonverbal gegeben werden können. Die soziale Beziehung zwischen Kind und Erwachsenen ist in viel stärkerem Maße asymmetrisch als in anderen Interviewsituationen, weil das Kind die Erwachsenen vor allem als Eltern und Lehrer versteht und nicht die Möglichkeit hat, sich selbst in die Rolle eines Interviewers zu versetzen. Für das Kind handelt es sich bei den Erwachsenen um Personen, die viel mehr wissen als Kinder und es daher eigentlich nicht nötig haben, Fragen zu stellen. Somit besteht die Gefahr, daß das Kind die Fragen als Prüfungsfragen mißversteht und unter Leistungsdruck nach Richtig-Falsch-Unterscheidungen sucht, statt nach Präferenzunterscheidungen. Das Kind weiß nicht, was man von ihm in der Interviewsituation erwartet, kennt also nicht die Rolle eines Befragten. Es empfiehlt sich daher, die unvertraute Forschungsaktivität in eine vertraute Situation zu verwandeln, indem ein Spiel als Rahmen für die Datenerhebung verwendet oder das Interview selbst als Spiel durchgeführt wird. So kann man zum Beispiel für die Gesprächsoperation Spielzeugtelefone benutzen oder Puppen als Interviewer einsetzen. Das Puppenspiel-Interview kann bei Kindern eine geeignete Datenerhebungsmethode sein, sofern die Einschränkungen, die für projektive Verfahren grundsätzlich gelten, bedacht werden. Das gilt gleichermaßen für die Verwendung von Bildvorlagen oder die Ergänzung unvollständiger Geschichten.

3. *Befragung im Handlungskontext*

3.1 Befragung und Introspektion

In der psychologischen Forschung werden Befragungen weniger zum Zwecke der Erhebung von Fakten- und Meinungsdaten vorgenommen als vielmehr in der Absicht, intra- und interindividuelle Differenzen von Informationsverarbeitungsprozessen zu ermitteln. Kognitionen sind der zentrale Gegenstand. Dabei kann man zwischen selbstbezogenen und umweltbezogenen Kognitionen unterscheiden, denen eine handlungsleitende Funktion zugeschrieben wird. Die Befragung erscheint nur unter der Voraussetzung sinnvoll, daß die handelnde Person über die eigenen Kognitionsinhalte Auskunft geben kann. Die Tatsache, daß man immer irgendwelche Daten erhält, wenn man Personen über ihre handlungsleitenden Kognitionen befragt, ist kein Nachweis für gültige Introspektionsvorgänge (Nisbett/Wilson 1977; Smith/Miller 1978). Offenbar ist die Befragung nach Beweggründen für routinisierte Handlungs- und Denkabläufe zwecklos, weil Introspektion in solchen Fällen zu nachträglich etablierten Kognitionen führt, die eine handlungskommentierende oder handlungsrechtfertigende Funktion haben. Die Unterscheidung von Kognitionen mit verschiedenen Funktionen im Handlungsverlauf stellt für den Forscher ein schwerwiegendes Problem dar. In einem Projekt zum Beispiel zur Erfassung von Lehrerkognitionen im Unterrichtsverlauf werden mit Hilfe von Videoausschnitten nachträglich Interviews durchgeführt, in denen eine Rekonstruktion von ehemals handlungsleitenden Kognitionen versucht wird (Wahl 1979). Dabei wird eine Widerspruchstechnik (Konfrontationsmethode) verwendet, wie sie in Streßinterviews üblich ist. Der Befragte wird in der verbalen Artikulation seiner vermeintlichen Kognitionen auf die Probe gestellt, er wird daran gehindert, auf der Basis von anfänglich geäußerten Kognitionsinhalten einfach weiter zu assoziieren.

Interviewvarianten, die auf diese Weise realisiert werden, sind halbstandardisiert und halbstrukturiert. Sie lassen sich den Intensivinterviews zuordnen, bei denen die Fragen offengehalten sind, um den Antwortspielraum zu vergrößern, und bei denen die Reihenfolge und Formulierung der Fragen auf den Befragten in der Situation selbst zugeschnitten sind (focused interview). Beim klinischen Interview wird diese Vorgehensweise ebenfalls bevorzugt, zum Beispiel wenn es um die subjektive Wahrnehmung der eigenen Lebensentwicklung und Konfliktbewältigungsversuche geht (life-history interview). Ein Interviewleitfaden genügt, um die erforderlichen Daten zu erheben. Eine andere Variante im klinischen Bereich ist das nondirektive Interview, welches in der Gesprächstherapie üblich ist. Es verfügt über keine vorgegebene Struktur, der Interviewer beschränkt sich auf akzeptierendes Kopfnicken und die Verbalisierung der emotionalen Erlebnisinhalte des Klienten mit dem Ziel, diesen bei der

Selbstexploration zu unterstützen. Die Datenerhebung ist dabei sekundär und dient lediglich der nachträglichen Evaluation des Therapeutenverhaltens.

Intensivinterviews zur Erfassung von handlungsleitenden Kognitionen sind sehr schwierig auszuwerten. Die theoretische Vorarbeit ist erheblich aufwendiger als zum Beispiel diejenige, die für die herkömmliche Umfrageforschung erforderlich ist. Die einzelnen Aussagen der Befragten müssen einem Netzwerk von Hypothesen zugeordnet und auf Konsistenz geprüft werden. Geht es dem Forscher um die Rekonstruktion von subjektiven Theorien als einer geordneten Menge von handlungsbezogenen Kognitionsinhalten, sind deren Strukturen zu erforschen. Scheele und Groeben (1979) haben dafür eine Struktur-Lege-Technik entwickelt, die als Grundlage für eine konsensorientierte Validitätsprüfung dienen kann. Bei dem Dialog-Konsens handelt es sich um eine gemeinsame Interpretation der Interviewdaten zwischen dem Forscher und dem Befragten. Der Befragte wird auf diese Weise als ein reflexives Subjekt angesehen und als Instanz zur Validierung der Daten genutzt. Der Befragte ist hier nicht Versuchsperson, sondern Versuchspartner.

Ein grundsätzliches Problem bei der introspektiven Datenerhebung liegt darin, daß man sich mehr an eigenes Handeln und Denken erinnert als an andere Situationsdeterminanten und sich selbst meist etwas positiver sieht, als es die neutralen Beobachter tun (self enhancement). Die aus dem Gedächtnis abgerufene Information ist in diesem Sinne selektiv (egocentric bias) und verfälscht die Daten (Ross/Sicoly 1979). Auf der anderen Seite kann gerade die Selbstbezogenheit im Denken und Handeln der Forschungsgegenstand sein. Die Analyse von Selbstgesprächen ist ein Weg zur Ermittlung der kognitiven Prozesse, die für bestimmte Formen des Erlebens und Handelns verantwortlich sind (Belschner 1980). Das Interview kann dabei die Funktion übernehmen, durch wenige gezielte Stimuli den Befragten bei der Rekonstruktion von selbstkommunikativen Abläufen zu unterstützen (lautes Denken). In der kognitiven Verhaltenstherapie spielt die Selbstinstruktion (inneres Sprechen) als Forschungsgegenstand und Therapieziel ebenfalls eine wichtige Rolle.

3.2 Intendierte Veränderungen im Forschungsprozeß

Die Befragung richtet sich im psychologischen Forschungsprozeß weniger auf vorgefundene Daten als auf solche, die während der Interaktionssituation erzeugt werden. Es handelt sich demnach um einen aktiven Konstruktionsprozeß von sprachlich abbildbarer Realität. Aus diesem Sachverhalt wird manchmal die Überlegung abgeleitet, den Interviewer aus seiner möglichst neutralen Rolle eines Instruments zu befreien und in besserer Nutzung seiner intellektuellen Flexibilität ihn bewußt als Mitgestalter solcher Realitätskonstruktionen einzusetzen. Damit werden Interviewer und Befragter zu Versuchspartnern,

deren Anliegen die gemeinsame Sinnkonstitution ist. Die Gültigkeitsprüfung erfolgt durch Konsensstechniken. Charakteristisch für den äußeren Ablauf solcher Interviews ist der Mangel an Strukturiertheit und Standardisiertheit. Statt dessen wird der Befragte lediglich dazu veranlaßt, sich im Hinblick auf eine Alltags- bzw. Problemsituation frei zu äußern (narratives Interview). Anstelle quantitativer Auswertungsmethoden werden interpretative Verfahren angewandt. Der wissenschaftliche Status solcher Befragungsmethoden ist umstritten. Sie sind grundsätzlich ungeeignet zum Gewinn generalisierbarer Erkenntnisse. Daher werden sie vor allem im Kontext von Handlungsforschung bevorzugt, in der die planmäßige Veränderung des Untersuchungsfeldes angestrebt wird.

Integrative Forschungs- und Handlungsprozesse in abgegrenzten Praxisfeldern können Befragungen als wichtige Elemente einschließen. Bei der Organisationsentwicklung und Systemberatung erweist es sich als sinnvoll, in einem ersten Schritt Daten über das System zu erheben. Das können zum Beispiel Einstellungs- und Konfliktbewältigungsmuster bei Lehrern und Schülern einer neuen Gesamtschule sein. In einem zweiten Schritt werden die Ergebnisse in das System zurückgemeldet (Survey feedback), um sie für die dort Handelnden transparent und diskutierbar zu machen. Anschließend werden strukturelle oder habituelle Veränderungen vorgenommen, die in einer letzten Phase empirisch evaluiert werden (Miles u.a. 1970). Der Unterschied zur herkömmlichen Umfrageforschung liegt vor allem in der Funktion der Befragung, die hier eine interventionsvorbereitende Methode darstellt.

Literatur

- Anger, H.: Befragung und Erhebung. 1969. In: C. F. Graumann (Hrsg.): Sozialpsychologie. Bd. 7/1 des Handbuchs der Psychologie. Göttingen: Hogrefe.
- Arntzen, F. 1970. Psychologie der Zeugenaussage. Göttingen: Hogrefe.
- Armstrong, J. S. 1975. Monetary incentives in mail surveys. *Public Opinion Quarterly*, 39, 111-116.
- Atteslander, P. 1969. Methoden der empirischen Sozialforschung. Berlin: de Gruyter.
- Babbie, E. R. 1973. *Survey Research Methods*. Belmont: Wadsworth.
- Bailey, K. D. 1978. *Methods of Social Research*. New York: Free Press.
- Baltes, P. B. (Hrsg.) 1979. *Entwicklungspsychologie der Lebensspanne*. Stuttgart: Klett-Cotta.
- Belschner, W. 1980. Konstruktion und Bearbeitung pädagogischer Situationen. In: W. Belschner, M. Doss, M. Hoffmann, F. Schott: *Verhaltenstherapie in Erziehung und Unterricht*. Band 2. Stuttgart: Kohlhammer, 99-168.

- Berdie, D. R. 1973. Questionnaire length and response rate. *Journal of Applied Psychology*, 58, 278-280.
- Binder, J., Sieber, M., Angst, J. 1979. Verzerrungen bei postalischen Befragungen: das Problem der Nichtantworter. *Zeitschrift für experimentelle und angewandte Psychologie*, 26, 53-71.
- Carpenter, E. H. 1975. Personalizing mail surveys: a replication and reassessment. *Public Opinion Quarterly*, 38, 614.
- Erbslöh, E. 1972. Interview. Stuttgart: Teubner.
- Erbslöh, E., Wiendieck, G.: Der Interviewer. In: J. van Koolwijk, M. Wieken-Mayser (Hrsg.) 1974. *Techniken der empirischen Sozialforschung*. Band 4: Die Befragung. München: Oldenbourg, 83-106.
- Esser, H.: Der Befragte. 1974. In: J. van Koolwijk, M. Wieken-Mayser (Hrsg.): *Techniken der empirischen Sozialforschung*. Band 4: Die Befragung. München: Oldenbourg, 107-145.
- Esser, H. 1977. Response set - methodische Problematik und soziologische Interpretation. *Zeitschrift für Soziologie*, 6, 253-262.
- Frey, D. (Hrsg.) 1978. *Kognitive Theorien der Sozialpsychologie*. Bern: Huber.
- Friedrichs, J. 1973. *Methoden empirischer Sozialforschung*. Reinbek: Rowohlt.
- Fisseni, H.-J. 1974. Zur Zuverlässigkeit von Interviews. *Archiv für Psychologie*, 126, 71-84.
- Fuller, C. 1974. Effect of anonymity on return rate and response bias in a mail survey. *Journal of Applied Psychology*, 59, 292-296.
- Heberleirr, T. A., Baumgartner, R. 1978. Factors affecting response rates to mailed questionnaires: a quantitative analysis of the published literature. *American Sociological Review*, 43, 447-462.
- Hensley, W. E. 1974. Increasing response rate by choice of postage stamps. *Public Opinion Quarterly*, 38, 280-283.
- Janke, W. 1973. Das Dilemma von Persönlichkeitsfragebogen. In: Reinert, G. (Hrsg.): Bericht über den 27. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1970. Göttingen: Hogrefe.
- Jöreskog, K. G., Sörbom, D. 1978. LISREL. Analysis of Linear Structural Relationships by the Method of Maximum Likelihood. Chicago: NER.
- Johnson, J. H., Sarason, I. G. 1979. Recent Developments in Research on Life Stress. In: V. Hamilton, D. M. Warburton (Eds.): *Human Stress and Cognition*. New York: Wiley, 205-233.
- Koolwijk, J. van. 1974. Die Befragungsmethode. In: J. van Koolwijk, M. Wieken-Mayser (Hrsg.): *Techniken der empirischen Sozialforschung*. Band 4. Die Befragung. München: Oldenbourg, 9-23.
- Kreutz, H. 1972. *Soziologie der empirischen Sozialforschung*. Stuttgart: Enke.
- Linsky, A. S. 1975. Stimulating responses to mailed questionnaires. *Public Opinion Quarterly*, 39, 82-101.

- Matteson, M. T. 1975. Type of transmittal letter and questionnaire colour as two variables influencing response rates in a mail survey. *Journal of Applied Psychology*, 59, 535-536.
- Miles, M. B., Hornstein, H. A., Callahan, D. M., Calder, P. H., Schiavo, R. S. 1970². The consequence of survey feedback: theory and evaluation. In: Bennis, W. G., Benne, K. D., Chin, R. (Eds.): *The planning of change*. New York: Holt, Rinehart & Winston, 457-468.
- Nisbett, R. E., Wilson, T. D. 1977. Telling more than we can know: verbal reports on mental processes. *Psychological Review*, 84, 231-259.
- Ross, M., Sicoly, F. 1979. Egocentric biases in availability and attribution. *Journal of Personality and Social Psychology*, 37, 322-336.
- Scheele, B., Groeben, N. 1979. Zur Rekonstruktion von subjektiven Theorien mittlerer Reichweite. Bericht Nr. 18 aus dem Psychologischen Institut der Universität Heidelberg. Heidelberg.
- Scherer, K. R. 1974. Beobachtungsverfahren zur Mikroanalyse non-verbaler Verhaltensweisen. In: J. van Koolwijk, M. Wieken-Mayser (Hrsg.): *Techniken der empirischen Sozialforschung*. Band 3: Beobachtung und Analyse von Kommunikation. München: Oldenbourg, 66-109.
- Scheuch, E. K. 1973³. Das Interview in der Sozialforschung. In: R. König (Hrsg.): *Handbuch der empirischen Sozialforschung*. Band 2. Stuttgart: Enke, 66-190.
- Schumann, H., Presser, S. 1978. The assessment of „No Opinion“ in attitude Surveys. In: K. F. Schuessler (Ed.): *Sociological Methodology 1979*. San Francisco: Jossey-Bass, 241-275.
- Schwarzer, R. 1975. Instrumente der empirischen Curriculumevaluation. In: K. Frey (Hrsg.): *Curriculum-Handbuch*. Band 2. München: Piper, 748-766.
- Scott, C. 1961. Research on Mail Surveys. *Journal of the Royal Statistical Society*, 124, 143-205.
- Smith, E. R., Miller, F. D. 1978. Limits on perception of cognitive processes: A reply to Nisbett and Wilson. *Psychological Review*, 85, 355-362.
- Sudman, S., Bradburn, N. 1974. *Response effects in surveys*. Chicago: Aldine.
- Suls, J. M., Miller, R. L. 1977. (Eds): *Social comparison processes*. Washington: Hemisphere.
- Triebe, K. 1976. *Das Interview im Kontext der Eignungsdiagnostik*. Bern: Huber.
- Wahl, D. 1979. Methodische Probleme bei der Erfassung handlungsleitender und handlungsrechtfertigender subjektiver psychologischer Theorien von Lehrern. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 11, 208-217.
- Wieken, K. 1974. Die schriftliche Befragung. In: J. van Koolwijk, M. Wieken-Mayser (Hrsg.): *Techniken der empirischen Sozialforschung*. Band 4: Befragung. München: Oldenbourg, 146-161.

7. Kapitel

Exploration

Udo Undeutsch

Die Exploration ist wahrscheinlich die umstrittenste psychologische Erhebungsmethode. Sie wird von einigen für die ergiebigste und sicherste Erkenntnisquelle gehalten, von anderen als wertlos eingeschätzt. Ebenso unterschiedlich ist der Begriffsgebrauch. Im englischen Sprachbereich wird der Begriff für Befragungstechniken überhaupt nicht verwendet, im deutschen Sprachbereich wird der Begriff in der Psychologie in der unterschiedlichsten Weise gebraucht: manchmal synonym mit Befragung überhaupt, manchmal eingeeengt auf ganz bestimmte Befragungstechniken, wobei wieder große Unterschiede zwischen verschiedenen Autoren bestehen.

1. Begriffsbestimmung

Ein *Gespräch* ist die partnerbezogene wechselseitige Ausübung der Sprechfähigkeit im zwischenmenschlichen Kontakt in der Absicht, einen Austausch von Innerlichkeitsgehalten vorzunehmen. Dabei wechseln die Rollen der Gesprächsteilnehmer zwischen Sprecher und Hörer. Diese Rollen können unter Gesprächsteilnehmern annähernd gleichgewichtig oder mit unterschiedlichem Gewicht verteilt sein. Sind die Rollen der Gesprächspartner stärker ungleichgewichtig verteilt, so ergeben sich daraus Sonderformen der Gesprächsführung: das Lehrgespräch, das therapeutische Gespräch, das Beratungsgespräch, das Verkaufsgespräch usw. auf der einen Seite und die verschiedenen Formen der Befragung auf der anderen Seite.

Die *Befragung* kann sehr unterschiedlichen Zwecken dienen und in sehr verschiedener Form vorgenommen werden.

1. Sie kann Wissen, Meinungen und Einstellungen über außerpersönliche Sachverhalte erkunden.
2. Sie kann der Vorbereitung einer helfenden (beratenden, therapeutischen usw.) Intervention dienen.

3. Sie kann den befragten Menschen selbst zum Ziel haben: sein Erleben und Verhalten in Vergangenheit und Gegenwart und sein „Wesen“.

Befragungen mit der erstgenannten Zielrichtung werden im deutschen Sprachgebrauch als *Interview* bezeichnet. Es ist in den empirischen Sozialwissenschaften das methodische Instrument mit der weitesten Verbreitung und der größten Zahl von Anwendungsmöglichkeiten. Methodologische Übersichtsreferate sind im deutschen Sprachbereich z.B. von Scheuch 1962, Anger 1969 und Erbslöh 1972 erstattet worden.

Für Befragungen der zweiten Art wird der Begriff *Anamnese* verwendet. Thoms (1975) definiert:

„Anamnese ist eine Methode klinischer Informationssammlung und bezeichnet gleichzeitig die ermittelten Daten. Gesprächsweise oder mit Fragebogen wird dabei die Vorgeschichte eines Menschen, bezogen auf eine bestimmte Fragestellung - psychische oder körperliche Symptomatik -, erhoben.“

Die Anamnese ist eine Methode der klinischen Psychologie. Sie ist auch keineswegs auf die Befragung des Patienten beschränkt. Fremdbeobachtungen und Dokumente werden regelmäßig einbezogen.

Befragungen mit der an dritter Stelle genannten Zielrichtung können sehr verschiedene Bindungsgrade aufweisen: von der vollkommen standardisierten Befragung über die teilstandardisierte Befragung bis hin zur nicht-standardisierten (= freien, ungebundenen) Befragung. Nur die letztgenannte Befragungsart wird im deutschen Sprachgebrauch als *Exploration* bezeichnet, wobei die Wahl eines Fremdwortes, das als terminus technicus gebraucht wird, anzeigt, daß nicht jede derartige Befragung, wie sie im Alltag in unzähligen Varianten vorkommt, gemeint ist, sondern nur die fachkundig vorgenommene psychologische (oder tiefenpsychologische oder psychiatrische) Befragung. Dies unterscheidet die Exploration vom Interview des Journalisten wie auch von der Vernehmung durch Polizei, Staatsanwaltschaft oder Gericht. Der Umfang der auf das Erleben und Verhalten des untersuchten Menschen und auf diesen selbst gerichteten Befragung kann wiederum sehr unterschiedlich sein: er reicht von einer Befragung, die „das Ganze eines individuellen Lebenslaufes erfassen soll“ (Thomae 1968, 113), bis hin zu speziellen Fragestellungen, wie es die Aufklärung eines vom Befragten erlebten Ereignisses ist. Es wäre mit dem fachwissenschaftlichen Sprachgebrauch im Deutschen nicht zu vereinbaren, den Ausdruck Exploration auf die das ganze bisherige Leben eines Menschen umfassende Befragung zu beschränken. Es ist aber zweckmäßig, thematisch begrenzte Befragungen, bei denen es nur einen bestimmten Lebensbereich abzuklären gilt, als „Exploration zur Sache“ (Undeutsch 1954, 15) zu spezifizieren. Entscheidend bleibt aber auch bei diesen stark themenzentrierten Befragungen, daß es sich um offene und wenig strukturierte Befragungen handelt, in denen der befragte Mensch sein Erleben berichtend ausbreiten

kann. *Exploration ist demnach die mit psychologischer Sachkunde vorgenommene nicht-standardisierte mündliche Befragung eines einzelnen Menschen durch einen einzelnen Gesprächsführer mit dem Ziel, Aufschluß zu erhalten über „Das Individuum und seine Welt“* (Thomae 1968). Sucht man nach einem etwa bedeutungsgleichen deutschen Wort, so kommt man schon vom genauen lateinischen Wortsinn her (ex-plorare = auskundschaften, erforschen) auf den Begriff „*Erkundungsgespräch*“, der auch schon von Arnold (1957) und Pongratz (1957) vorgeschlagen worden ist.

2. Geschichte

Es läßt sich die Annahme begründen, daß das Erkundungsgespräch im Alltagsleben so alt ist wie der menschliche Sprachgebrauch selbst. Die Frage des Diomedes an seinen Gegner auf dem Schlachtfeld

„Wer doch bist Du, Edler, der sterblichen Erdenbewohner?“ (Homer: *Ilias*, 6. Ges., Z. 123)

(die damals dazu führte, daß die Gegner sich als Freunde aus Väterzeiten erkannten und ihre Freundschaft erneut beschlossen, statt gegeneinander zu kämpfen) ist in dieser und ähnlicher Form (freilich nicht immer mit dem damaligen erfreulichen Erfolg) eine der Grundfragen des menschlichen Alltags. Als terminus technicus ist der Begriff „Exploration“ in der klassischen Psychiatrie entstanden, wo darunter das Eruiere psychopathologischer Phänomene beim Patienten verstanden wurde. Der Begriff wurde in weiterer Bedeutung in die Psychologie übernommen von Binet und Piaget.

Die ersten thematischen Ansätze zur Erforschung von Individuen und ihrer Welten durch Befragung und andere biographischen Methoden finden sich bei W. Stern (1900, 3. Aufl. 1921) und seinen Mitarbeitern (Baade und Lipmann 1909, Margis 1911). In der wissenschaftlichen Forschung ist die Erhebungsmethode der Exploration intensiv angewandt worden etwa innerhalb der Child-Guidance-Untersuchung von MacFarlane (1938), in den Jahren 1938 bis 1947 von Kinsey und seinen Mitarbeitern (1948 und 1953) zur Erforschung des sexuellen Verhaltens des Menschen. Bezüglich des Einsatzes der Exploration in der Persönlichkeitsforschung verdient Pfahler (1939) der Vergessenheit entrissen zu werden. Sodann ist aber vor allem auf Thomae zu verweisen, der durch die Art des Einsatzes der Exploration und der Auswertung von Explorationsbefunden gerade jene Gebiete der Persönlichkeitserforschung erschlossen hat, die bisher von der Wissenschaft beiseite gelassen worden waren, nämlich das „alltägliche“ wie auch das „krisenhafte“ Verhalten des Menschen in „natürlichen“ Situationen (1968). Nach Thomae bildet die Exploration

„einen der wenigen Zugänge zu einer durch den methodischen Zugriff noch nicht veränderten seelischen Wirklichkeit“ (1968, 113).

Sein Hauptwerk „Das Individuum und seine Welt“ (1968) kann man nach seinen eigenen Worten

„als einen Beitrag zur Technik der Auswertung von systematisch gewonnenen Explorationsprotokollen und von Protokollen über Verhaltensbeobachtungen aus unterschiedlich langen biographischen Einheiten ansehen“ (117).

Für charakterdiagnostische Zwecke hat die Exploration eine zentrale Rolle gespielt in der deutschen Wehrmachtpsychologie (1927-1945; s. hierzu Kreipe 1936, Walther 1941, Beck 1942, Kröber 1942, Mierke 1944, 66-70). In abgewandelter Form wurde die Exploration von der US-amerikanischen Militärpsychologie übernommen (Assessment 1948). Seit Herbst 1944 wurde die Exploration zusätzlich zu einer Serie von Fähigkeitstests und einem Persönlichkeitsfragebogen bei der Auswahl von Militärpiloten der schwedischen Luftwaffe (Trankell 1956) und seit 1951 bei der Auswahl der Piloten der SAS (Trankell 1959) verwendet.

Der Wert der wissenschaftlich ausgestalteten Exploration für die Beurteilung der Eignung von Bewerbern wurde natürlich auch für Berufe der freien Wirtschaft entdeckt. In England hatte schon vor dem Zweiten Weltkrieg Oldfield im Auftrag des National Institute for Industrial Psychology die Erfahrungen aus Einstellungsgesprächen gesammelt und ausgewertet und daraus eine Methodik des Einstellungsgesprächs entwickelt (1951). Parkinson (1957) schrieb eine geistreiche Parodie über das in England bei Behörden wie Industrieunternehmen übliche Bewerbungsgespräch. Eine noch größere Rolle spielte das Bewerberinterview (employment interview, selection interview) im US-amerikanischen Wirtschaftsleben, weil dort - wegen des geringen Ausleseeffekts des dortigen Schul- und Bildungssystems - die Bewerber viel weniger durch Schulbildung und Berufsausbildung vorsortiert sind, weil eine viel größere Fluktuation der Arbeitskräfte besteht und weil, vor allem in den unteren Bevölkerungsschichten, Ehrlichkeit und Redlichkeit gegenüber dem Arbeitgeber sehr zu wünschen übrig lassen. Nach einer Erhebung von Spriegel und James (1958) gaben im Jahre 1930 93% von 236 befragten Firmen an, daß sie ihre Bewerber vor der Einstellung einem Interview unterziehen. Bellows und Estep schätzten 1954 die Zahl der jährlich in den USA zur Bewerberauslese durchgeführten Interviews auf 150 Millionen. Dort ist deshalb schon seit Jahrzehnten eine reiche Literatur über Explorationstechnik in Monographien (Kephart 1952, Bellows & Estep 1954, Fear 1953, 2. Aufl. 1978, The McGraw Hill course in effective interviewing, 1973) sowie in Handbüchern der Industriepsychologie (Roethlisberger und Dickson 1939, Bellows 1949, 3. Aufl. 1961) vorhanden.

Eine andere Entwicklungslinie leitet sich her aus der sehr viel älteren Kriminaltaktik, die mit Vernehmungpsychologie (H. Gross 1893, 1898) und Aussagepsychologie (W. Stern 1902) wichtige Beiträge zur Explorationstechnik ver-

heimlicher Sachverhalte und zur Bewertung von Explorationsdaten geliefert hat. Auf die Bedeutung der „experimentell gestalteten“ „Exploration zur Sache“ hat Undeutsch für die Aussagepsychologie (1954, 15 und 1967, 117) und für die Beurteilung der Schuldfähigkeit (1965, ebenso Thomae und Schmidt 1967, 354-356) hingewiesen. Ebenso hat Undeutsch zu wiederholten Malen auf die Bedeutung der Exploration in der Fahrereignungsdiagnostik (für die Abschätzung der Rückfallwahrscheinlichkeit bei Verkehrsdelinquenten unter Alkoholeinfluß sowie bei mehrfach durch Verstöße in nüchternem Zustand auffällig gewordenen Verkehrsdelinquenten) hingewiesen.

Welche Bedeutung die Exploration im methodischen Arsenal der Psychologie heute erlangt hat, kann man am besten daran erkennen, daß im „Handbuch der Psychologie“ in den bisher erschienenen Bänden keine andere Methode der Datengewinnung so häufig behandelt worden ist wie sie. Es sind ihr vier umfassende Artikel gewidmet worden:

In ihrer allgemeinsten Form, aber ausschließlich im Hinblick auf ihre Verwendung in den Sozialwissenschaften, behandelt die Befragung der Artikel von

Anger: Befragung und Erhebung. 1969. 7/1, 567-617.

Die Bedeutung der Exploration in der Persönlichkeitsdiagnostik behandelt der Artikel von

Schraml: Das psychodiagnostische Gespräch. (Exploration und Anamnese). 1964, 6, 868-897.

Ihre Bedeutung in der Form der Anamnese für die Klinische Psychologie der Artikel von

Kemmler und Echelmeyer: Anamnese-Erhebung. 1978. 8/2, 1628-1648.

Ihre Bedeutung für die Forensische Psychologie der Artikel von

Friedrichs: Die aussagepsychologische Exploration. 1967, 11, 3-25.

Die Zahl der empirischen Untersuchungen zur Methode der Exploration ist rund um die Welt in den letzten Jahren gewaltig angewachsen.

3. Qualitative Charakterisierung

Die Exploration gehört zu den qualitativen Methoden par excellence. Deren methodische Eigentümlichkeiten sind herausgearbeitet z.B. in den Artikeln von W. Salber (1960, 1969). Die dort herausgestellten methodischen Grundzüge der qualitativen Methoden haben samt und sonders auch für alle explorativen Verfahren Geltung.

Im Sinne der Methodensystematik von Cattell (1957) gehört die Exploration zu den Verfahren, welche der Gewinnung von „Life-data“ dienen. Sie ist die Hauptmethode einer „psychologischen Biographik“, die nach Thomae als eine Synthese zwischen „ideographischer“ und „nomothetischer“ Forschung aufgefaßt werden kann und deren Ziel es ist, „eine psychologische Analyse des menschlichen Verhaltens im natürlichen Ablauf des Lebens zu erarbeiten“ (1968, 105).

Die Exploration *unterscheidet* sich von anderen Methoden der Datenerhebung.

Sie unterscheidet sich vom *Experiment* dadurch, daß das von ihr erfaßte Erleben und Verhalten der Untersuchungssituation vorausliegt und daher Bestandteil des natürlichen Ablaufs des Verhaltens des Individuums bleibt, während das Experiment in die Lebensumstände des Untersuchten eingreift, bestimmte Bedingungen herstellt und diese planmäßig verändert. Sie kann aber andererseits von der experimentellen Methodik profitieren, indem sie die Bedingungen der Befragung planmäßig verändert, was alsdann zu der Beobachtung Gelegenheit gibt, welche Veränderungen dadurch in den Mitteilungen der Befragten hervorgerufen werden (Undeutsch 1954, 15). Diese Veränderungen können zur Beurteilung des Wahrheitsgehalts der gemachten Mitteilungen herangezogen werden.

Sie unterscheidet sich vom *Test* dadurch, daß sie sich nicht darauf beschränkt, nur einen durch standardisierte Testbedingungen definierten ganz bestimmten Ausschnitt aus dem Verhaltensrepertoire in Anspruch zu nehmen. In Tests werden in der Regel Auslösereize dargeboten, die dem Untersuchten entweder kaum oder jedenfalls weit weniger vertraut sind als dem Untersucher, während in der Exploration typischerweise eine „Gemeinsamkeit der Erlebnissphäre“ (Thomae 1968, 112) besteht.

Zudem ist die Testsituation im allgemeinen durch eine gewisse „Neutralität“ und Konstanz der Situation gekennzeichnet. Die in den hohen Testwiederholungskoeffizienten zutage tretende hohe Stabilität und Konstanz des Reagierens der Individuen sind eher Ausdruck einer Versuchstechnik, welcher es gelingt, aus allen „existentiell“ begründeten Varianten des Verhaltens, wie sie sich schon im Laufe eines Tages, erst recht aber im Laufe eines Lebens ergeben, auf eine relativ neutrale Sphäre auszuweichen. Das Ziel der persönlichkeitsdiagnostischen Verfahren, nämlich bestimmte Persönlichkeitszüge mit konstanter Zuverlässigkeit zu erfassen, wird in den Tests geradezu dadurch erreicht, daß diese gegen den Aufforderungscharakter existentiell bedeutsamer Aspekte der jeweiligen Situation immun gemacht werden. Als Folge davon haben diese diagnostischen Verfahren, welche von dieser Konzeption der „existentiellen Neutralität“ ausgehen, für die Erfassung des realen individuellen Verhaltens und seiner „dispositionellen“ Hintergründe nur bedingt Wert. In

der Exploration ist es demgegenüber möglich, das Verhalten des Individuums in der Vielfalt der banalen wie der existentiell bedeutsamen Situation seines realen Lebensvollzugs zu erfassen.

Auch der standardisierte *Fragebogen* läßt der persönlichkeitspezifischen Aktivitätsentfaltung keinen Spielraum. Die erwähnte „Gemeinsamkeit der Erlebnissphäre“ ist auch bei den Fragebögen stärker beeinträchtigt, da diese so konstruiert sind, daß vom Befragten ein hoher Prozentsatz der gestellten Fragen verneinend beantwortet werden muß.

„Die Vermutung, eine ganze Serie von ‚neurotischen Symptomen‘ zu haben, wirkt auf ‚Normale‘ nicht gerade kontaktstiftend“ (Thomae, ebd.).

Thomae führt zusammenfassend aus:

„Die Gewinnung dieser Verhaltens-Daten hat dem ideographischen Prinzip zu folgen, das möglichst die ‚unverzerrte‘ psychische Wirklichkeit zu erfahren und zu erfassen strebt. Deshalb können Tests und Fragebögen nicht die *primäre* Quelle dieser Forschung sein, da sie in jedem Falle ja eine Veränderung des realen Verhaltens (Umschaltung von einer sinnbezogenen ‚erfüllten‘ Situation auf eine relativ sinnarme, nur durch Zusatzmotivationen zu stabilisierende) darstellen“ (1968, 106).

Am schwierigsten ist naturgemäß die Abgrenzung der Exploration gegen das *Interview*. Thomae (1968; 112) stellt zunächst die Gemeinsamkeiten heraus: Beide stellen eine Form der verbalen Kommunikation dar. Gemeinsam ist beiden weiterhin, daß Auslösereize (z.B. Fragen) verwendet werden, die einer beiden Partnern in ähnlicher Weise vertrauten Erlebnissphäre angehören. Dennoch besteht zwischen Interview und Exploration ein Unterschied, den es mit Wellek (1958) und Thomae (1968, 112) festzuhalten gilt. Nicht nur ist das Interview in seinen verschiedenen Formen mehr oder minder standardisiert, sondern es ist auch ausschließlich oder doch vorwiegend auf Wissen, Meinungen, Vorstellungen von außerhalb des befragten Subjekts gelegenen „Dingen“ ausgerichtet. Demgegenüber ist es für die Exploration typisch, daß sie auf den befragten Menschen selbst gerichtet ist, der nicht nur als Reflektor, sondern in seiner Eigenschaft als Subjekt, als Mitmensch, als Partner Gegenstand des ganzheitlichen explorativen Interesses ist.

Von allen anderen Datenerhebungsverfahren unterscheidet sich die Exploration dadurch, daß sie nicht wie diese die Antwortmöglichkeiten des Untersuchten auf ein Konzept einengt, das den Erwartungen einer bestimmten Theorie oder den Erfordernissen einer bestimmten Methodologie entspricht. Dadurch bleibt den anderen Verfahren der Zugang zur vollen Breite menschlichen Verhaltens verschlossen.

„Da eine Fremdbeobachtung dieses Verhaltens aus äußeren Gründen meist nicht möglich ist, stellt die Exploration einen der wenigen Zugänge zu einer durch den methodologischen Zugriff noch nicht veränderten seelischen Wirklichkeit dar“ (Thomae 1968, 113).

Für die Erforschung der Existenzweise und der Funktion bestimmter „inhaltlicher“ Strukturen des „subjektiven Lebensraumes“, des Selbst und der am meisten relevanten „thematischen“ und „instrumentellen“ Einheiten innerhalb der menschlichen Existenz ist die Exploration unentbehrlich und unersetzbar, weil die untersuchten Phänomene dort untersucht werden müssen,

„wo sie aller Voraussicht nach anzutreffen sind, nämlich im ‚alltäglichen‘ wie u.U. auch im ‚krisenhaften‘ Verhalten des Menschen. Dieses in der erforderlichen Extensität und Intensität von fremden Beobachtern aus zu erfassen, wird jedoch durch hoffentlich kaum überwindbare ethische Vorentscheidungen verwehrt. In dieser Situation muß die Wissenschaft versuchen, dem einzig verfügbaren Zeugen, nämlich dem Subjekt des Geschehens, mehr Glauben zu schenken. Seine Aussagen, die wir durch die Exploration zu erfassen und zu vertiefen versuchen, stellen von hier aus gesehen die einzige sichere Quelle für die Erschließung des Verhaltens in ‚natürlichen‘ Situationen dar“ (Thomae 1968, 222f.).

Das schließt nicht aus, daß die Exploration in bestimmten ihrer Formen (Anamnese) und in bestimmten Phasen (gegen Ende) eine gewisse Standardisierung erfahren kann. Auch die Exploration kann systematisch auf die relevanten Gesichtspunkte der jeweils vorliegenden Fragestellung eingehen, wobei auch eine weitgehende Standardisierung der einzelnen Fragen, unter Umständen auch der Reihenfolge, vorgenommen werden kann. Prinzipiell müssen alle Fragen aber „offene Fragen“ bleiben.

Auch eine Quantifizierung von Explorationsergebnissen ist keineswegs ausgeschlossen. So kann z.B. im Verlauf einer Exploration der bisherigen Verkehrsgeschichte allmählich die Zahl der Verkehrsunfälle erfragt werden, an denen der Befragte beteiligt war (v. Klebelsberg 1970, S. 45), oder es kann zur Prognose der Rückfallwahrscheinlichkeit in eine erneute Fahrt in alkoholbedingt fahruntüchtigem Zustand die größte im letzten Jahr vor der Untersuchung aufgenommene Menge an alkoholischen Getränken exploriert werden (Welzel 1976). Außer solchen trivialen Quantifizierungsmöglichkeiten bestehen zahlreiche Skalierungsmöglichkeiten von formalen Grundkategorien des in der Exploration erfaßbaren Verhaltens eines Menschen in den natürlichen zeitlichen Einheiten seines Lebens. Bei einem von Thomae unternommenen Versuch, das von verschiedenen Menschen geschilderte Geschehen unter Abstraktion von seinem jeweiligen Gehalt deskriptiv zu erfassen, ergaben sich mehrere formale Aspekte des biographischen Geschehens, die auf die natürlichen Einheiten des individuellen Bios - der Handlung, des Tageslaufes und der mehr oder minder großen faßbaren Ausschnitte des Lebenslaufes (z.B. Formen der Auseinandersetzung mit einer beruflichen oder familiären Situation) - anwendbar sind. Die Skalen können, je nach dem Ausmaß der vorhandenen Informationen, wenige (z.B. 5) oder viele (z.B. 9) Stufen haben (Thomae 1968, 124-216).

4. Methodische Prinzipien des explorativen Gesprächs

Ziel der in der Form des Gesprächs angestellten Erkundungsbemühungen ist das konkrete mitmenschliche „Individuum und seine Welt“. Dieses menschliche Individuum wird veranlaßt, Auskunft über sich und „seine“ persönlich-individuelle „Welt“ zu geben, über das, was es erlebt hat und der Erinnerung und der Erwähnung für wert hält, über seinen „psychologischen Lebensraum“ (K. Lewin 1936, S. 14ff.; im deutschen Originaltext 1969, 40-50). Das ist „alles, was vom Standpunkt des Psychologen für diese Person existiert“ (40), der

„Inbegriff dessen, was jeweils für das betreffende Individuum wirksam ist. Für die Aufgabe der begrifflichen Ableitung des Geschehens kann man Wirksamkeit als das Kriterium für psychologische Existenz verwenden: ‚*Wirklich ist, was wirkt*‘“ (41).

Dies erfordert die uneingeschränkte Kooperationsbereitschaft des Befragten. Seine Bereitschaft muß vom Explorator gewonnen werden. Alle explorations-technischen Anleitungen beschäftigen sich daher damit, wie diese grundlegende Voraussetzung am besten herzustellen sei.

Bedingungen für das Zustandekommen dieser Bereitschaft werden zunächst in der *Persönlichkeit des Explorators* gesucht. Verlangt werden: persönliche Reife, Reichtum an Vorstellungs- und Miterlebensmöglichkeiten (Rowe 1963, Blakeney & McNaughton 1971) und an Lebenserfahrung, Kontaktfähigkeit und charakterliche Werte wie Vertrauenswürdigkeit (für alles Vorstehende: Taft 1955, Wiley & Jenkins 1964, Steinkamp 1966) und eine Haltung dem Mitmenschen gegenüber, die C. G. Jung vor vielen Jahren (1932) einmal wie folgt gekennzeichnet hat:

„Will der Arzt die Seele eines Anderen führen, oder sie auch nur begleiten, so muß er mit ihr Fühlung haben. Diese Fühlung kommt nie zustande, wenn der Arzt verurteilt, ob er das nun mit so viel Worten laut tut, oder unausgesprochen im Stillen, ändert nichts an der Wirkung. Auch das Umgekehrte, nämlich dem Patienten unbesehen Recht geben, hilft nichts, es wirkt ebenso entfremdend wie das Verurteilen. Fühlung entsteht nur durch vorurteilslose Objektivität. Das klingt beinahe wie etwas Wissenschaftliches. Man könnte es mit einer rein intellektuellen, abstrakten Einstellung verwechseln. Was ich aber meine, ist etwas ganz anderes: Es ist etwas Menschliches, etwas wie eine Hochachtung vor der Tatsache, vor dem Menschen, der an dieser Tatsache leidet, vor dem Rätsel eines solchen Menschenlebens“ (Ges. W. 1963, 11, 366f.).

Darüber hinaus stellt ein *sozialpsychologischer Bedingungs-zusammenhang* eine grundsätzlich wichtige Voraussetzung für das Zustandekommen dieser Bereitschaft dar: Jeder Mensch teilt sich lieber und leichter gegenüber einem anderen Menschen mit, wenn er überzeugt sein darf, daß der andere ihn versteht, denn nicht verstanden zu werden, birgt in sich fast unvermeidlich die Gefahr, falsch und daher ungerecht beurteilt zu werden.

„A subject is inevitably hesitant to discuss things which seem to be both outside of the experience of the interviewer, and beyond his knowledge“ (Kinsey u.a., 1948, 60).

So ist es denn auch gewiß kein Zufall, daß z.B. die ergiebigsten Explorationen mit Homosexuellen von homosexuellen Wissenschaftlern geführt worden sind. Das gleiche ist natürlich auch bei anderen Personengruppen zutage getreten: z.B. bei Alkoholabhängigen, Drogensüchtigen oder auch bei Ordensleuten. Cospers (1969) untersuchte die Möglichkeit des Interviewer-Bias systematisch, indem er das Trinkverhalten seiner 28 Interviewer sehr detailliert erhoben hat. Alle Interviewer waren Alkoholkonsumenten und wurden ausführlich geschult, um das Interview möglichst zu standardisieren. Trotzdem konnten eindeutige Effekte nachgewiesen werden: stärker trinkende Interviewer erhalten höhere Quantitätsangaben zum Trinken; die Angaben zur Häufigkeit des Konsums variieren nicht mit dem Trinken der Interviewer. Dies ist ein Hinweis darauf, daß Quantitätsangaben zum Trinken schwieriger zu erhalten sind als Häufigkeitsangaben. Es läßt sich recht genau angeben, worauf sich das Vertrauen auf die Verständnisfähigkeit und -bereitschaft beim Explorierten zu gründen pflegt:

1. Eine (vermutete, angenommene, erlebte) - mindestens partielle - Wesensverwandtschaft zwischen Explorator und Exploriertem. So hat sich z.B. in Untersuchungen von Ledvinka (1971, 1972) die Auswirkung von Rassen-gleichheit bzw. -Verschiedenheit auf die vom Explorierten gemachten Mitteilungen gezeigt.
2. Eine (vermutete, angenommene, erlebte) aus - mindestens partieller - gleichartiger Lebenserfahrung oder wenigstens
3. aus Sachkunde und Berufserfahrung erwachsene Vertrautheit mit dem explorierten Lebensgebiet.

Die Vertrautheit mit dem explorierten Lebensgebiet muß der Explorator gegenüber dem Explorierten während der Exploration unter Beweis stellen:

1. durch sachkundige, intime Kenntnis des betreffenden Lebensgebietes verratende Fragerichtung,
2. durch Nachweis seiner Vertrautheit mit typischen Zusammenhängen zwischen einzelnen Gegebenheiten des betreffenden Lebensgebietes,
3. durch insider-typischen Wortgebrauch,
4. durch Vermeidung aller Äußerungen des Erstaunens, der Überraschung, der positiven oder negativen Bewertung des vom Explorierten berichteten Verhaltens oder Erlebens.

Der Explorator muß mit der „Welt“, in der der Explorierte lebt, vertraut sein: mit seinem äußeren Lebensraum, mit seiner sozialen Schicht und seinem Berufsstand, mit seinen Lebensbedingungen, mit seiner Sprache und seinen Wertvorstellungen. Er muß aber auch vertraut sein mit dem Lebensgebiet oder

dem Sachgebiet, dem die Exploration gilt. Es genügt auch nicht, daß der Explorator solche Vertrautheit behauptet, sondern diese muß während der Exploration, in deren Gestaltung und ihrem Verlauf, in Erscheinung treten, muß für den Explorierten auf Schritt und Tritt spürbar werden, damit die „Gemeinsamkeit der angesprochenen Erlebnissphäre“ (Thomae 1968, 112) zum Tragen kommen kann. Kinsey u.a. (1948) gibt dafür ein eindrucksvolles Beispiel:

„A single phrase from an understanding interviewer is often sufficient to make the subject understand this, and such an interviewer wins a record where none would have been disclosed to the uneducated investigator. A specific illustration will make this more apparent.

This is the case of the older Negro male whose first answers were wary and evasive. When questioned concerning his occupation, he listed a variety of minor jobs which, taken in connection with his manner of response, seemed to spell underworld activities. We followed up our clue by immediately asking the subject whether he had ever been married. We were not satisfied with his denial of marriage, and followed with a question as to whether he had ever lived common law. The easy use of a vernacular term made him feel freer to talk, and when he admitted that he had so lived, we asked how old he was when he first lived common law. When he said that he was then fourteen, our first suspicion concerning his underworld activity was confirmed, and we immediately followed up by asking how old the woman was. At this, he smiled and admitted that she was thirty-five. Then we remarked, easily and without surprise: „She was a hustler, wasn't she?“ This was the final step necessary for winning complete confidence. The subject stopped short in his reply, opened his eyes wide, smiled in a friendly fashion, and said, ‚Well, sir, since you appear to know something about these things, I'll tell you straight.‘ The extraordinary record that we then got of his history as a pimp could not have been obtained if the subject had not comprehended that we understood the world in which he lived“ (60f.).

Das erfordert vom Explorator viel Einarbeitung und Vorbereitung. Als Beispiel sei erwähnt, daß Kinsey und seine Mitarbeiter vor Beginn ihrer allein aus Explorationen gewonnenen Erhebung mehrere Jahre darauf verwendet haben, sich die erforderlichen Sachkenntnisse auf dem von ihnen erforschten Lebensgebiet anzueignen, und ein weiteres volles Jahr, um sich in die Explorationstechnik einzüben (1948, 61).

„Very often the interviewer's capacity to secure an accurate history depends upon his knowledge of the correlations that usually exist between certain items, and his readiness to demand an explanation of any inconsistency that appears in a particular history. To illustrate again: one starts by asking the girl how old she was when she turned her first trick (but one does not ask how old she was when she was first paid as a prostitute). She is then asked how many of the tricks return after their first contacts with her. Considerably later in the interview there is a question concerning the frequency with which she rolls her tricks (robs her customers). The girl who reports that few of the men ever return, and who subsequently says that she never robs any of the men, needs to be caught up abruptly and assured that you know that it doesn't work that way. If she

doesn't roll any of the men, why don't they return to her? This question is likely to bring a smile from the girl and an admission that since you appear to know how these things work, she will tell you the whole story, which means that she robs every time there is any possibility of successfully doing so" (Kinsey u.a., 1948, S. 61).

Zu den genannten persönlichen und sozialpsychologischen Voraussetzungen muß auf seiten des Explorierten hinzukommen die *Bereitschaft, sich mitzuteilen*. Diese Bereitschaft muß zu Beginn der Exploration geweckt und während des gesamten Verlaufes unterhalten werden durch die gewissermaßen werbepsychologischen Maßnahmen (B. Spiegel 1965) des Informierens und des Motivierens: des Informierens über den Zweck der Untersuchung und des Motivierens zu aufrichtigen, ehrlichen, wahrheitsgemäßen und vollständigen Angaben. Die konkrete Ausgestaltung des Informierens und des Motivierens im Zuschnitt auf den konkreten Einzelfall wird je nach dem Untersuchungszusammenhang und je nach den Besonderheiten des Einzelfalles sehr verschieden aussehen können und müssen. Hier sind die Unterschiede sehr groß zwischen der Gewinnung der Mitarbeit für eine wissenschaftliche Untersuchung auf der einen Seite und der sehr viel schwerer zu gewinnenden Mitarbeit für eine Untersuchung, deren Ergebnis nicht von vornherein und zwangsläufig dem Untersuchten zugute kommt. Handelt es sich um die Mitarbeit an einer wissenschaftlichen Untersuchung, so sind oftmals schon die Befriedigung darüber, vor einem verständigen Zuhörer über sich selbst sprechen zu können, und die Genugtuung darüber, an einem anerkannt wichtigen Unternehmen mitzuwirken, eine ausreichende und im allgemeinen leicht zu erreichende Motivationsbasis (Kinsey 1948, 36; Whyte 1955; Bain 1960; Thomae 1968, 114). Sehr viel schwieriger ist die Motivation zu wahrheitsgemäßer und vollständiger Auskunft zu erzielen, wenn der Befragte - zu Recht oder zu Unrecht - gerade davon unerwünschte oder sogar direkt nachteilige Konsequenzen für sich zu befürchten hat. Das ist nicht nur der Fall bei allen Bewerberexplorationen, sondern in noch viel höherem Maße bei Fahrereignungsuntersuchungen, bei denen der Untersuchte selbstverständlich die Entdeckung seiner Ungeeignetheit mit allen Mitteln zu verhindern trachtet, und erst recht bei allen Untersuchungen im gerichtlichen Auftrag - sei es z.B. bei der Sorgerechtszuteilung oder bei der Rekonstruktion des Herganges eines Schadensereignisses zum Zwecke der Schadenersatzregelung im zivilrechtlichen Bereich oder bei der Rekonstruktion der Entstehung und des Ablaufes einer tatbestandsmäßig rechtswidrigen Handlung, der Motive und der geistig-seelischen Verfassung des Täters dabei im strafrechtlichen Bereich. Dennoch lehrt die Erfahrung, daß es auch in solchen Situationen - sogar in der Regel - möglich ist, eine aufrichtige Kooperationsbereitschaft des Befragten zu erzielen, wenn der Fragende psychologisch geschult ist und die vernehmungpsychologischen Richtlinien (Inbau & Reid 1974) richtig anwendet:

„Since he is trained in his task, he will typically sympathize with the suspect, provide face-saving rationalizations for any crimes that might have been committed, and indi-

cate subtly that he can understand how someone might murder but that lying is the lowest form of degeneracy: Such a procedure may sound farfetched, yet in the hands of an expert it is remarkably convincing" (Orne, Thackray, and Paskewitz 1972, 746).

Die konkrete Ausgestaltung des Informierens und Motivierens ist für die einzelnen Untersuchungsanlässe und für die einzelnen Personen zu verschieden, um im Rahmen dieses Artikels dargestellt werden zu können. Vortreffliche Anleitungen dazu sind in der Literatur leicht zugänglich. Weithin bekannte Anleitungen zur Explorationstechnik sind: Richardson, Dohrenwend, and Klein, 1965; Bingham, and Moore, 1. ed. 1931, 4. ed. with the collaboration of J. W. Gustad, 1959; Kahn & Cannell 1957 (10. printing 1966); Fear 1953, sec. ed. 1978. Weniger bekannt ist, daß es gerade für die Exploration schwieriger, heikler Sachverhalte, bezüglich deren beim Befragten starke Tendenzen zur Verheimlichung bestehen, sehr gute und ganz aufs Praktische ausgerichtete Anleitungen gibt: Kinsey, Pomeroy & Martin 1948, 35-62, deutsche Ausgabe 1964, 22-52; Inbau & Reid 1974.

Als Voraussetzung für die Objektivität und die Differenziertheit der Auswertung der Explorationsbefunde ist es erforderlich, Aufnahmen der Exploration auf Tonträgern und maschinenschriftliche Übertragungen anzufertigen. Froehlich (1958) verglich für 97 anamnestiche Explorationen, die zur Vorbereitung von psychologischen Beratungen durchgeführt worden waren, die Aufzeichnungen der therapeutischen Berater mit den Tonbandaufnahmen dieser Explorationen. Seine Untersuchung kam zu dem Ergebnis, daß weniger als ein Drittel der Inhalte im schriftlichen Bericht auftauchte, diese aber mit 75% bis 94% Korrektheit. Es ließ sich jedoch keine klare Beziehung zwischen der Wichtigkeit der Daten und ihrer Wiedergabe durch den Therapeuten im schriftlichen Bericht erkennen. Froehlich fand in seiner Erhebung keinen signifikanten Unterschied zwischen erfahrenen und weniger erfahrenen Beratern. Zu einem gleichartigen Ergebnis kommt Thomae (1968):

„Als Ergebnis langer Erfahrungen in der Handhabung der explorativen biographischen Anamnese muß aber leider darauf verwiesen werden, daß Gedächtnisprotokolle für eine systematische Auswertung völlig unzureichend sind und das Mitschreiben bzw. -stenographieren erhöhte Anforderungen an den Untersucher stellt und das Gespräch sehr stört. Auch hierbei kann unter Auswertung aller früher genannten Motivationen die zunächst unmöglich erscheinende Zustimmung erreicht werden. In der Regel wird bei einem gut geführten Gespräch das mitlaufende Tonband vergessen“ (115).

Durch die vorstehenden Ausführungen ist auch

„die in Deutschland und überhaupt im kontinentalen Europa vorherrschende Ansicht, Exploration sei eine Kunst, welche man nicht lehren und lernen und damit auch nicht empirisch untersuchen könne“ (Schraml 1964, 869),

widerlegt. Alle praktischen, künstlerischen und wissenschaftlichen Fertigkeiten sind lehrbar und lernbar, wenn der Erfolg natürlich auch je nach den

mitgebrachten Persönlichkeits- und Begabungseigenschaften unterschiedlich sein wird. Das ist für die Explorationstechnik nicht anders als für die Ausbildung in ausübender Musik oder in Mathematik, ohne daß den Wissenschaftscharakter mathematischer Methoden jemals einer aus diesem Grund zu bestreiten sich veranlaßt gesehen hätte (W. Metzger 1942). Hinweise zur Didaktik der Ausbildung finden sich u.a. bei Schraml (1964, 889), Rechetnick & Barkus (1966), Soudijn u.a. (1970), Thorne (1970), Wolfe (1970), Meier (1972), Schuller & Rosemeier (1973), Froehlich & Bishop (1973). Von erfolgreicher Ausbildung für die Durchführung von Explorationen berichtet Kinsey (1948, 61 f.). Ausbildungserfolge in der Auswertung von biographischen Explorationen beweist ein Versuch von U. Lehr (1964), der S. 26 näher geschildert wird. Es waren Explorationsausschnitte, die Monate oder Jahre später gewonnen worden waren, Ausschnitten aus einer Exploration der gleichen Personen zu einem früheren Zeitpunkt und zu anderen Themen zuzuordnen. Die Trefferquote betrug bei 30 Studenten des ersten Semesters der Psychologie 67,6%, bei den in psychologischer Diagnostik ausgebildeten Hauptdiplom-Kandidaten dagegen 835%.

5. Auswertung

Die methodischen Einwände gegen eine verbreitete Anwendung der Exploration ergeben sich im wesentlichen aus Bedenken gegen den Wahrheitsgehalt (die Ehrlichkeit und die objektive Richtigkeit) der Mitteilungen. Es handelt sich dabei vor allem um zwei prinzipielle Fehlerquellen: Verfälschung und Verheimlichung. Der Tatbestand der Verfälschung liegt vor, wenn der Befragte zwar richtige Angaben machen könnte, sich aber entschließt, stattdessen eine bewußt entstellte Darstellung zu geben, oder etwas tatsächlich Nichtvorhandenes zu behaupten, während der Tatbestand der Verheimlichung vorliegt, wenn er sich entschließt, den erfragten Sachverhalt ganz oder teilweise wahrheitswidrig in Abrede zu stellen. Die Tendenzen zur Verfälschung und zur Verheimlichung spielen naturgemäß eine außerordentlich unterschiedliche Rolle, je nachdem ob die Exploration zu einem neutralen wissenschaftlichen Zweck, bei zugesicherter Anonymität bei der Verwertung der Explorationsergebnisse, geschieht oder in einer Situation des realen Lebens, in der für den Untersuchten elementare vitale Interessen auf dem Spiele stehen. Möglichen Verfälschungs- und Verheimlichungstendenzen muß schon bei der Erhebung der Explorationsbefunde mit explorationstechnischen Maßnahmen entgegengewirkt werden. Dennoch ist es in Situationen, die dem Befragten zu einer absichtlichen Verfälschung oder Verheimlichung Anlaß geben können, erforderlich, den gegebenen Bericht an Hand der in der Aussagepsychologie erarbeiteten Kriterien für den Wahrheitsgehalt von Aussagen (Undeutsch 1967) zu überprüfen (Thomae und Schmidt 1967, 354; Böcher 1968; Kunkel 1978, 96-110). Eine weitere Überprüfungsmöglichkeit besteht im Vergleich der

Mitteilungen mit den Ergebnissen empirischer Forschungen auf dem betreffenden Lebensgebiet. Ein Beispiel dafür war oben schon mit Kinseys Hinweisen für die Exploration von weiblichen Prostituierten („Beischlafdiebstahl“) gegeben worden. Die Überprüfung des Wahrheitsgehaltes der Mitteilungen durch den Vergleich mit den Ergebnissen empirischer Forschungen ist nicht nur während der Exploration erfolgreich einzusetzen, sondern dient auch nachträglich zur Überprüfung des Wahrheitsgehaltes der Explorationsbefunde. Darauf hat ebenfalls schon Kinsey (1948) hingewiesen:

„When 90 to 95 per cent of the persons *in* any social level report histories which agree with the patterns shown in Chapter 10, they not only establish the nature of the group Patterns, but establish the validity of their own reports as well“ (129).

Andere Beispiele gibt Kunkel (1978, 65-89, 99-102, 104f.): Die Behauptung eines Kraftfahrers z.B., die abgeurteilte Trunkenheitsfahrt sei die erste seines Lebens gewesen, ist an sich schon unwahrscheinlich wegen der hohen Dunkelziffer von Trunkenheitsfahrten (1:400). Sie wird noch unwahrscheinlicher, wenn man in Rechnung zieht, daß die BAK sehr hoch war und daß er nichtsdestoweniger eine längere Strecke unauffällig gefahren war und auch selbst bekundet, er habe sich voll fahrtüchtig gefühlt. Da auch das Fahren unter Alkoholeinfluß gelerntes Verhalten ist, muß ein Fahrer bereits häufiger unter Alkoholeinfluß gefahren sein, bis es ihm gelingt, eine längere Strecke mit einer so hohen BAK unauffällig zurückzulegen. Beispiele für diese Art der Überprüfung von Explorationsangaben lassen sich in großer Zahl aus fast allen Lebensbereichen finden.

Für die Auswertung von Explorationsdaten für die persönlichkeitspsychologische Forschung hat Thomae (1968) eine umfassende und größtenteils bahnbrechende Anleitung gegeben. über die oben bereits erwähnte Einstufung des explorativ erhobenen biographischen Geschehens unter mehreren formalen Gesichtspunkten (1968, 124-216) hinaus hat Thomae Listen von Dimensionen entwickelt zur Kennzeichnung der inhaltlichen Aspekte (unter Gesichtspunkten der Expansion, des strukturellen Aufbaus und der qualitativen Dimensionen) des subjektiven Lebensraumes (S. 223-256. Tab. 17-20) und des Selbst (S. 256-282, Tab. 22-28). Er hat weiter gezeigt, daß explorativ gewonnenes biographisches Material sieben fundamentalen thematischen Einheiten zugeordnet werden kann (292-328) und daß aus diesem Material die instrumentellen Einheiten oder Aspekte des personalen Geschehens, die er als „Daseinstechniken“ bezeichnet, herausgearbeitet werden können. „Daseinstechniken“ sind die persönlichkeitsspezifischen Arten und Weisen, wie sich das Individuum sein Leben innerlich wie äußerlich „möglich“ bzw. „erträglich“ zu machen sucht. Thomae hat über 20 Typen oder Klassen solcher instrumentellen Einheiten herausgearbeitet, die den verschiedenen Ebenen einer biographischen Analyse (Tagesläufen und den größeren biographischen Abschnitten) gemeinsam sind, und diese zu 5 Typen oder Grundklassen funda-

mentaler instrumenteller Einheiten zusammengefaßt (349-366). Bei Anwendung dieser Auswertungskategorien kann das in Explorationen gewonnene Material wichtige und einzigartige Beiträge zur Entwicklungsforschung, zur Persönlichkeitsforschung und zur Individualdiagnose der untersuchten Menschen liefern.

Die Verarbeitung der Explorationsbefunde zu einer Diagnose oder Prognose setzt weiter voraus, daß der diagnostische oder prognostische Wert des Befundes - für sich allein oder in Verbindung mit anderen Befunden - aus empirischen Untersuchungen bekannt ist. Ein mustergültiges Beispiel für die Verarbeitung der prognostischen Bedeutsamkeit von biographischen Daten und ihrer Gewichtung für die Erstellung einer Prognose hat Kunkel für die „Rückfallprognose bei Trunkenheitstätern im Straßenverkehr“ (1977) gegeben.

6. Leistungsfähigkeit der explorativen Methoden

Für die Bemessung der Leistungsfähigkeit messender diagnostischer Verfahren wird in der klassischen Testtheorie unterschieden zwischen der Leistungsfähigkeit des betreffenden Verfahrens als Meßinstrument (Meßgenauigkeit in bezug auf eine gegebene Population = Reliabilität) und seiner Leistungsfähigkeit für den vorgesehenen Zweck (Vorhersagegenauigkeit, Tauglichkeit = Validität). Die übliche Behandlung der Objektivität als eines weiteren Gütekriteriums (z.B. Lienert 1969, 13) ist gewiß verfehlt; Objektivität ist nur eine Komponente der Reliabilität, genauer: eine von mehreren Arten reliabilitätssenkender Fehler (Ekman 1947, 158f.).

Es gibt eine unübersehbar große Zahl von Arbeiten, in denen diese Testparameter für einzelne explorative Techniken oder für einzelne Explorationsdaten berechnet worden sind. Es bedarf zuvor aber einer kritischen Besinnung darauf, ob die Leistungsfähigkeit eines so andersartigen methodischen Instruments, wie es die explorativen Techniken darstellen, durch diese Maßwerte überhaupt sinnvoll gekennzeichnet werden kann. Es darf nicht aus dem Auge verloren werden, daß die klassische Testtheorie, der die Begriffe Reliabilität und Validität entstammen, zur Bestimmung der Leistungsfähigkeit von messenden Verfahren entwickelt worden ist, d.h. von Verfahren, die in ihrer Durchführung und Auswertung vollständig oder weitgehend standardisiert sind, die in einer vom „natürlichen“ Leben abgehobenen, künstlichen, gleichartig und konstant gehaltenen Situation durchgeführt werden und die sich darauf beschränken, einen mehr oder minder klar umgrenzten Persönlichkeitsbereich oder sogar nur eine spezielle Begabungskomponente zu erfassen. Grundlegend für die klassische Testtheorie ist die Voraussetzung der Verfügbarkeit „äquivalenter“, d.h. vergleichbarer Messungen derselben Eigenschaft. In allen genannten Hinsichten sind die explorativen Techniken nicht nur an-

ders, sondern geradezu entgegengesetzt beschaffen. Schon die projektiven diagnostischen Verfahren haben sich für eine Bemessung ihrer diagnostischen Leistungsfähigkeit nach den Kriterien der klassischen Testtheorie als weitgehend unzugänglich erwiesen. Die unter Anwendung dieser testtheoretischen Verfahren bei den projektiven Tests erzielten Ergebnisse sind bekanntlich allesamt sehr unbefriedigend geblieben. Der Grund dafür ist ein doppelter: Zum einen ist eine Berechnung der Kennzahlen der klassischen Testtheorie nur sinnvoll, wenn die Befähigung des Psychologen zur Durchführung und zur Auswertung dieser Verfahren von völlig untergeordneter Bedeutung ist, so daß das Verfahren von jedem Psychologen „gleich gut“ angewandt werden kann - oder es für die Durchführung und die Auswertung überhaupt nicht einmal eines Psychologen bedarf. Wenn nämlich zwischen den Psychologen erhebliche persönliche Unterschiede hinsichtlich ihrer Qualifikation (auf Grund von Begabung, Ausbildung, Übung und Erfahrung) für die Durchführung und die Auswertung eines Verfahrens bestehen, so kann jede Maßzahl für die Leistungsfähigkeit eines Verfahrens nur einen Durchschnittswert aus den Leistungen von Testanwendern der allerverschiedensten Qualitätsstufen darstellen. Es besagt aber selbstverständlich nichts gegen den Wert eines Verfahrens, wenn sich bei Überprüfungen seiner Leistungsfähigkeit herausstellt, daß einige damit hervorragend zu arbeiten verstehen, während viele andere nur sehr mäßige Erfolge damit erzielen und einige andere sogar zu vorwiegend falschen Ergebnissen damit gelangen. Der Validitätskoeffizient ist in solchen Fällen nur ein nichtssagender Mittelwert - nichtssagend, weil er die viel belangreichere Tatsache überdeckt, daß es Testanwender gibt, die mit dem Test hervorragende Ergebnisse zu erzielen verstehen, während es gleichzeitig andere gibt, in deren Händen die Testanwendung lediglich Unfug ist. Ein Beispiel dafür ist die Untersuchung von Magnusson (1959) zur Validität des TAT:

An 63 männlichen Studenten aus einem Hochschulinternat wurde der TAT durchgeführt. Die Testergebnisse wurden von 4 Psychologen ausgewertet, die ihre Kompetenz für TAT-Auswertungen erklärt hatten. Die gleichen Studenten wurden von je 10 bis 18 Kommilitonen, die seit 2 Jahren im Internat mit ihnen zusammenlebten, in 19 Persönlichkeitsvariablen eingestuft. Sowohl die Psychologen als auch die Kommilitonen hatten für die Einstufung bei allen 19 Variablen eine 7stufige Skala zu verwenden. Für jedes Individuum wurde für jede Persönlichkeitsvariable der Mittelwert der Einstufungen durch seine Kommilitonen berechnet. Diese Mittelwerte bildeten die Kriterienvariable. Die Einstufungen der 4 Psychologen wurden mit den Kriterienvariablen korreliert. Diese Korrelationskoeffizienten sind Validitätskoeffizienten.

Die Validitätskoeffizienten fielen für die 4 Psychologen sehr unterschiedlich aus. Bei dem Psychologen B trat kein einziger negativer Korrelationskoeffizient auf, von seinen positiven Korrelationen waren 8 auf dem 10%-Niveau oder einem noch höheren Niveau signifikant. Der Psychologe A hatte hingegen ebenso viele positive wie negative Koeffizienten. Die Verteilung der von ihm erreichten Validitätskoeffizienten entspricht genau einer Zufallsverteilung (89-91, 105-107).

Für diagnostische Verfahren dieser Art gilt der Satz von Vernon (1957):

„... it is hardly possible to dissociate the test from the tester. One clinical psychologist does well with interviewing or with Rorschach, another with Thematic Apperception or drawings, another with deterioration tests or expressive movements and so on“ (205).

Ein anderes sehr eindrucksvolles Beispiel ist die Beurteilung des intellektuellen Niveaus mit Hilfe der graphologischen Diagnostik.

In einer Untersuchung von Michel (1969), an der 7 Diplom-Psychologen, die zusätzlich eine gründliche graphologische Ausbildung genossen hatten und über ausreichende praktische Erfahrung auf diesem Gebiet verfügten, teilgenommen hatten, betrug der Korrelationskoeffizient zwischen den graphologischen Schätzungen des Intelligenzniveaus und den Gesamtstandardwerten des IST, die als Kriterienvariable genommen worden sind, beim besten Graphologen + .57, beim schlechtesten - .29. Dieser Befund ist interessant, denn er besagt, daß der eine Graphologe recht gut zur Einschätzung des intellektuellen Niveaus auf Grund der Handschrift in der Lage war, während am anderen Ende der Skala ein Graphologe steht, dessen Urteil in der Mehrzahl der Fälle in der verkehrten Richtung lag. Was besagt demgegenüber der durchschnittliche Validitätskoeffizient von .16? Er verdeckt das wahre Ergebnis der Untersuchung.

Die Frage nach der Reliabilität und der Validität eines Verfahrens ist überhaupt nur sinnvoll, wenn es sich um ein standardisiertes Verfahren handelt, während die explorativen Techniken qualitativ (dem Modus nach) eine außerordentliche Spielbreite haben. Ein Verfahren, das von Mensch zu Mensch von Mal zu Mal wechselt, kann einer statistischen Überprüfung im klassischen Sinne überhaupt nicht unterzogen werden, weil es keine feste und gleichbleibende Gestalt hat. „Every interview is a unique and unreproducible encounter“ (Lopez, 1965, 8). Auch ist bei keinem anderen diagnostischen Verfahren der Verwendungszweck so vielfältig. So vielfältig wie die Verwendungszwecke, sind natürlich auch die Validitätskoeffizienten. Selbst für den Teilbereich der eigensdiagnostischen Exploration betonen Bolton & Hikey (1969):

„Interviews are not generally predictive; that is, they are not generally valid. Rather their validity must be determined in a given situation, for particular positions, and following specified procedures“ (501).

Ein anderer Grund für die Schwierigkeit der Anwendung der Maßstäbe der klassischen Testtheorie auf projektive Verfahren ist, daß Voraussetzung jeder Validitätsbestimmung ist, daß eine vom zu prüfenden diagnostischen Verfahren unabhängige, ihrerseits reliable und mit dem angezielten psychologischen Konstrukt möglichst identische oder wenigstens hochkorrelierende Kriterienvariable zur Verfügung steht. Daran gebricht es schon bei projektiven Tests in aller Regel, denn je komplexer das diagnostische Verfahren ist, um so komplexer und weniger klar umgrenzt ist der Komplex der psychologischen Konstrukte, die damit erfaßt werden können, wobei das gleiche diagnostische

Verfahren bei verschiedenen Menschen sogar verschiedene Ausschnitte aus dem Gesamt ihrer personalen Struktur erfassen kann, um so unmöglicher, für den Komplex der erfaßten überdauernden personalen Sachverhalte eine Kriterienvariable zu finden oder einen für alle untersuchten Individuen gleichen Komplex von solchen. Ein Vergleich mit Einzelkriterien kann die Leistungsfähigkeit der Exploration prinzipiell nicht in Erscheinung treten lassen. Mit Recht betont Wellek (1958):

„Ein ‚ganzheitliches‘ Verfahren wie die (echte) Exploration kann der Natur der Sache nach nur überlegen, ja sogar nur brauchbar sein, wenn das Erkenntnisziel - der gefragte Gegenstand - gleichfalls ein ganzheitliches ist“ (27).

Daraus wird deutlich, daß die Maßstäbe der klassischen Testtheorie auf Verfahren, die einerseits Ansprüche an die Qualifikation des Untersuchers und Auswerters stellen und andererseits selbst sehr komplex, vielgestaltig, von Fall zu Fall wechselnd sind, entweder überhaupt nicht oder allenfalls unter vielen Vorbehalten und mit nur sehr beschränkter Aussagekraft anwendbar sind.

Gilt das alles schon ganz allgemein für alle in der Durchführung und Auswertung nur wenig standardisierten Verfahren, so gilt es in höchstem Maße für die Exploration, die einer weitgehenden Standardisierung nicht unterworfen werden kann, ohne gleichzeitig denaturiert zu werden. Man muß daher gerade im Falle der Exploration nach anderen Bewährungskriterien Ausschau halten. Maßwerte für Reliabilität und Validität können nur randständige Bedeutung haben und sind wegen der Vielgestaltigkeit der explorativen Techniken und der Vielzahl der Verwendungszwecke ohnehin nur für den speziellen Fall, für den sie ermittelt worden sind, aussagefähig.

6.1 Reliabilität

Für die Bestimmung der Reliabilität ist Voraussetzung, daß „äquivalente“, d.h. vergleichbare Messungen derselben Eigenschaft zur Verfügung stehen. Es stehen vier empirische Methoden zu ihrer Bestimmung zur Verfügung: die Testwiederholung, der Paralleltestvergleich, die Testaufspaltung und die Populationsaufspaltung nach Husen (1949, 62-70, 75), wovon die letztgenannte im Fach unbekannt geblieben ist und deshalb auch keine Anwendung erfahren hat. Alle diese Methoden sind auf explorative Verfahren nicht anwendbar.

Die Testwiederholung verlangt, daß derselbe Test zu einem späteren Zeitpunkt nochmals unter genau den gleichen Bedingungen angewandt wird. Der gleichen wäre nur annäherungsweise bei einem streng standardisierten Interview, aber nicht bei einer aus dem unmittelbaren zwischenmenschlichen Kontakt lebenden Exploration möglich, weil jede Exploration in einem lebendigen Austausch von Fragen und Mitteilungen besteht. Kein Mensch antwortet aber

wie eine Grammophonplatte in der zweiten Exploration genauso wie in der ersten. Jede Änderung der Art und Weise des Antwortens hat aber eine Änderung der Fragenzahl und der Fragenformulierung des Explorators zur Folge. Deshalb kann niemals der gleiche Satz von Fragen in der gleichen Reihenfolge und in der gleichen Formulierung unter gleichen Bedingungen wiederholt werden.

Die Paralleltestmethode setzt voraus, daß ein Paralleltest vorhanden ist, der aus äquivalenten Items besteht. Auf die Exploration angewandt, müßte entweder der gleiche Explorator zwei Sätze äquivalenter Fragen haben oder der gleiche Fragensatz müßte von zwei Exploratoren den gleichen Personen vorgelegt werden. Das erste geht nicht, weil es überhaupt keine Methode gibt zur Überprüfung der Parallelität der beiden Fragensätze. Das zweite geht nicht, weil die gleichen Personen nicht annähernd zum gleichen Zeitpunkt von einem anderen Explorator in der gleichen Weise (hinsichtlich des Zwecks, der Dauer, der Zahl und der Art der behandelten Themen, der Reihenfolge der Fragen, der Fragenformulierung, der Art des zwischenmenschlichen Kontaktes usw.) exploriert werden können. Vielmehr werden verschiedene Exploratoren unvermeidlich persönlichkeitspezifisch verschieden explorieren und die Explorierten auf verschiedene Exploratoren verschieden reagieren, wie schon Nietzsche wußte (Menschliches, Allzumenschliches, Aph. 374) und hernach Klages (1926) weiter ausgeführt hat:

„Mit wieviel Menschen einer in Berührung zu kommen pflegt, über ebensoviel verschiedene Physiognomien verfügt seine Seele. Wir wollen uns an Beispielen verständlich machen. - Die Art und Weise, wie man auf eine und dieselbe Frage Antwort gibt, hängt ganz wesentlich von demjenigen ab, der die Frage stellt. Einen und denselben Vorgang erzählen wir diesem Zuhörer nicht mit genau den nämlichen Worten als jenem“ (19).

Die Exploration des einen ist naturnotwendig anders als die des anderen. Es liegen also gerade keine „Parallelen Tests“ vor, d.h. Tests, die sich gleichen „wie ein Ei dem anderen“ - ganz abgesehen davon, daß sich die testtheoretischen Kennwerte für Parallelität von Tests

$$(M_{X_1} = M_{X_2}, s_{x_1}^2 = s_{x_2}^2, r_{x_1x_2} = r_{x_3x_4}, r_{X_1Y} = r_{X_2Y})$$

für Explorationen jedes Sinnes entbehren und sich übrigens auch gar nicht berechnen lassen.

Es hat nicht an Versuchen gefehlt, die Vielzahl der reliabilitätssenkenden Fehler, die bei den explorativen Methoden eine Rolle spielen können, in einer komplizierten Formel zu vereinigen, um auf diese Weise trotz der hohen Komplexität der explorativen Methoden doch noch die Reliabilität berechnen zu können. Für Fälle, in denen die Antworten der Befragten quantifiziert

werden können durch Zuordnung einer Maßzahl (= Rohwert), sieht Fleiss (1970) die Rohwerte als zusammengesetzt an aus folgenden Parametern

- μ = eine additive Konstante,
- α = die Auswirkung des personspezifischen Explorationsstils des Explorators auf die Antworten des Explorierten und dadurch auf die Bewertung der Antworten,
- β = die Kriterien, die der Explorator üblicherweise bei der Bewertung der Antworten anwendet,
- γ = die resultierende Wirkung des Kompromisses, den der Explorator zu schließen hat zwischen den Anforderungen der Exploration und den Anforderungen der Bewertung der Antworten (für eine genaue Auswertung der Exploration ist es wichtig, daß die Antworten des Explorierten vollständig mitgeschrieben werden. Der Exploration hingegen bekommen die durch das Nachschreiben bedingten längeren Sprechpausen schlecht),
- d = die Zufallsverteilung der individuellen Eigenarten der Explorierten,
- e = ein Zufallseinfluß bei der Quantifizierung

und folgenden Interaktionsausdrücken:

- αd = die Interaktion zwischen Explorator und Exploriertem, die sich einerseits in der Weise ausdrücken kann, daß der Explorator im Hinblick auf die Besonderheiten des Explorierten von seinem habituellen Explorationsstil abweicht, und andererseits in der Weise, daß der Explorierte selbst bei gleicher Fragestellung gegenüber einem Explorator anders antwortet als beim anderen,
- βd = die Interaktion, derzufolge der Explorator möglicherweise bei verschiedenen Explorierten verschiedene Bewertungskriterien anwendet,
- γd = die unterschiedliche Gestaltung des Kompromisses im Hinblick auf verschiedene Explorierte.

Daraus ergibt sich, daß der im Falle einer Quantifizierung eines Explorationsbefundes gegebene Wert X sich wie folgt zusammensetzt

$$X = \mu + \alpha + \beta + \gamma + d + (\alpha d) + (\beta d) + (\gamma d) + e.$$

Es ist offensichtlich, daß die einzelnen testtheoretischen Parameter und Interaktionsausdrücke für die konkreten Einzelfälle nur schwer mit Zahlen ausgefüllt werden könnten. Und wenn dies möglich wäre, wäre eine solche Berechnung auch nur für explorative Techniken möglich, die in Zahlenwerte (scores) ausmünden. Und selbst bei solchen explorativen Techniken wäre der dann gewonnene Zahlenwert von höchst fragwürdiger Bedeutung, denn er vereinigt die Leistungen „guter“ und „schlechter“ Explorationsführer und die Antworten ergiebiger und unergiebigender, selbständiger und beeinflusbarer, aufrichtiger und unaufrichtiger usw. Explorationsteilnehmer. Es ist darum auch nicht verwunderlich, wenn in Sammelreferaten zum Wert der Exploration für die Berufseignungsdiagnostik (Wagner 1949, Ulrich und Trumbo 1965) immer wieder festgestellt wird, daß nur in wenigen Untersuchungen Reliabilitätskoeffizienten mitgeteilt werden. Bei den von Wagner referierten 106 Untersuchun-

gen enthielten nur 25 überhaupt quantitative Angaben. Die mitgeteilten Reliabilitätskoeffizienten lagen in Fällen der Beurteilung umschriebener Merkmale zwischen .23 und .97, für eine pauschale Eignungsbeurteilung zwischen .20 und .85. Mayfield (1964) kommt in einem Sammelreferat, ebenfalls über die Bedeutung der Exploration bei der Bewerberauslese für berufliche Positionen, zu dem Ergebnis, daß allgemeine Beurteilungen der Eignung eines Bewerbers auf der Basis unstrukturierter Explorationen und ohne weitere Vorinformationen eine extrem geringe Inter-Rater-Reliabilität aufwiesen und daß die gleichen Explorationsdaten von verschiedenen Beurteilern in unterschiedlicher, in einigen Fällen sogar in entgegengesetzter Weise interpretiert und gewichtet werden. Dieses Ergebnis sollte indessen nicht als Hinweis auf eine geringe Reliabilität der Exploration genommen werden, sowenig wie es üblich ist, die Tatsache, daß verschiedene Kliniker bei der Beurteilung von MMPI-Profilen nur geringe Übereinstimmung zeigen, gegen die Reliabilität des MMPI als Untersuchungsinstrument ins Feld zu führen. Denn natürlich kann man auch die mit anerkanntermaßen hochreliablen Meßverfahren gewonnenen Befunde diagnostisch falsch verwerten.

Es gibt andere methodische Möglichkeiten, sich von der Reliabilität (wenn man schon bei diesem Begriff bleiben will) der Exploration ein Bild zu machen. U. Lehr (1964) hat einen Versuch unternommen, der als eine der Natur der Exploration gemäße Abwandlung der Testwiederholungsmethode angesehen werden kann.

Das Versuchsmaterial bestand aus den Explorationen von 3 weiblichen Personen. Diese Personen

- waren zur Zeit der Exploration alle gleich alt: 29 Jahre;
- alle hatten die gleiche Schule besucht und im gleichen Jahr Abitur gemacht.
- Ebenso erhielten alle drei eine eingehende Berufsausbildung, gaben jedoch mit dem Zeitpunkt der Heirat die Berufstätigkeit auf;
- alle hatten Kinder.

Aus den Explorationen dieser 3 Frauen wurden je 3 Ausschnitte maschinenschriftlich übertragen, die sich jeweils bezogen

- auf die frühe Kindheit bis zum ersten Schultag,
- auf die Situation der Berufswahl,
- auf die Situation der Partnerwahl.

Die zur gleichen Person gehörenden Explorationsausschnitte waren mit der gleichen Signatur kenntlich gemacht.

Außerdem erhielten die Versuchspersonen die maschinenschriftlichen Übertragungen von Äußerungen der gleichen Personen, die einige Wochen, zum Teil ein Jahr später erhoben worden waren als der Lebenslauf. Diese Äußerungen bezogen sich

- auf soziale Einstellungen,
- auf die Bewertung des bisherigen Lebens und auf die Zukunftsorientierung.

Diese Äußerungen waren nicht signiert.

Diese Materialien wurden einer Gruppe von **30** diagnostisch geschulten Hauptdiplom-Kandidaten übergeben mit der Aufgabe, diese Ausschnitte jenen 3 Personen zuzuordnen, von denen sie Ausschnitte aus der früheren Exploration erhalten hatten. Sowohl die Auszüge aus den Explorationen zum Lebenslauf als auch die später erhobenen Äußerungen waren von allen Informationen gereinigt worden, die als Hinweise auf die Identität der Personen hätten dienen können. Von den 630 vorgenommenen Zuordnungen waren **515** = 83,5% richtig.

Die in der Exploration ersichtlich werdende Verhaltensgestalt kann somit als ein Spiegelbild bestimmter Verhaltenstendenzen, spezifischer Grade, Formen und Richtungen der Aktivität, bestimmter Interessenbevorzugungen und Neigungen angesehen werden. Da der Gegenstand der später aufgenommenen Äußerungen völlig von dem der Exploration des Lebenslaufes abwich, manifestiert sich offensichtlich in den Explorationsausschnitten unabhängig von dem konkreten Inhalt eine durchgängige persönlichkeitspezifische Ausprägung dieser Tendenzen so eindeutig, daß eine Zuordnung von solchen Äußerungen zu anderen, die zeitlich und inhaltlich von jenen deutlichen Abstand zeigen, möglich ist. Thomae (1968) interpretiert dieses Versuchsergebnis folgendermaßen:

„Die Exploration liefert also, wenn sie kunstgerecht und in einer für den Explorierten akzeptablen Motivationslage durchgeführt wurde, kein ad hoc und kein bewußt oder willkürlich zurechtgemachtes Material. Sie ‚entfaltet‘ vielmehr zumindest einen gewissen Ausschnitt aus den Verhaltensweisen des Individuums, den Situationen, so wie sie das Individuum erlebt, und den Zielsetzungen, von denen diese Verhaltensweisen her verstanden werden müssen“ (119).

Sehr gründlich hat sich bezüglich der Zuverlässigkeit der durch Explorationen erhobenen Daten Kinsey vergewissert. Er hatte den einzigartigen Vorteil, ein sehr großes Explorationsmaterial zur Verfügung zu haben (im Laufe der Jahre 1938 bis 1947 hatten er und seine 3 Mitarbeiter 12214 Explorationen durchgeführt), das zu mancherlei Kontrolluntersuchungen Gelegenheit gab.

a) 162 Personen wurden nach Ablauf einer längeren Zeit (18 Monate bis 7 Jahre) einer Zweitbefragung unterzogen. Die Ergebnisse sind in Tab. 13 (122f.) zusammengestellt. Es zeigte sich, daß eine hohe Wiederholungsübereinstimmung bestand hinsichtlich der Angaben darüber, ob eine bestimmte Form sexueller Aktivität vom Befragten jemals ausgeübt worden war. Die Wiederholungskoeffizienten lagen für alle diesbezüglichen Fragen über .90 und in allen bis auf 3 Fälle über .95. Hohe Wiederholungsübereinstimmung besteht ebenfalls für die mehr äußeren biographischen Daten, bei denen die Korrelationskoeffizienten in jedem Fall über .80 lagen, in 6 von 8 Fällen über .90. Deutlich niedriger waren hingegen die Angaben über das Alter bei der ersten Erfahrung mit einzelnen Formen sexueller Betätigung ($r = .5 - .8$). Dennoch sind die Differenzen nicht groß. Sie betragen im allgemeinen 5% oder weniger des arithmetischen Mittels. Am niedrigsten

sind die Koeffizienten bei denjenigen Erlebnissen, die keinen abgrenzbaren Ereignischarakter haben: kindliche sexuelle Spieltätigkeit, unwillkürliche nächtliche Ejakulationen, heterosexuelles Petting. Die Angaben über die durchschnittliche Häufigkeit, mit der die einzelnen Betätigungsformen während eines bestimmten Zeitraumes ausgeübt worden sind, korrelierten miteinander zwischen .58 und .67.

- b) Eine weitere Überprüfungsmöglichkeit ergibt sich aus dem Vergleich der Angaben von Ehepartnern. Kinsey hat solche Vergleiche für 231 Ehepaare vorgenommen (Tab. 14, S. 126). Verglichen wurden deren Angaben zu insgesamt 32 Einzelpunkten. Bei 3/4 der erfragten Einzelpunkte liegen die Übereinstimmungskoeffizienten über .70, bei der Hälfte über .80 und bei 1/4 über .90. In der Hälfte der Einzelpunkte besteht zwischen den Angaben der beiden Ehepartner eine nahezu vollständige Übereinstimmung in 90 bis 100% der Berichte. Diese Übereinstimmung ist erstaunlich, wenn man in Betracht zieht, daß zwischen der Befragung der beiden Ehepartner Zeiträume von 2 bis 6 Jahren (und mehr) lagen.
- c) Die Explorationen für den Kinsey-Bericht sind überwiegend von 3 Wissenschaftlern ausgeführt worden. Das gab die Möglichkeit, die Frage zu untersuchen, ob verschiedene Exploratoren gleiche Resultate zu erreichen pflegen. Es wurden zu diesem Zweck verglichen die Ergebnisse, die sie bei gleichartigen (parallelen) Gruppen von Befragten gewonnen hatten. Die Gruppen waren homogen unter den Gesichtspunkten des Geschlechts, der Rasse, des Familienstandes, des Alters, des Bildungsgrades, der Stadt-/Land-Zugehörigkeit, der Konfession. Es wurden Vergleiche durchgeführt für die Gruppen, in denen jeder Untersucher mehr als 300 Fälle exploriert hatte. In Tab. 16 (S. 134) sind 75 Vergleichswerte wiedergegeben. 35 davon
- „are so similar that the differences are immaterial - closer than any person could calculate about his own history“ (135),

wobei wiederum die Zahlen über die Häufigkeit, mit der die einzelnen Formen sexueller Betätigung praktiziert worden sind, eine etwas geringere Übereinstimmung aufwiesen als die Zahlen über die Verbreitung. Das Ergebnis dieser Vergleiche ist:

„There seems no reason to doubt that any other group of investigators could duplicate these results if their scientific objectivity and their methods in interviewing were comparable to those used in the present study“ (135).

- d) Kinsey, der persönlich im Laufe eines knappen Jahrzehnts 7036 Explorationen durchgeführt hatte, hatte noch eine weitere Möglichkeit zur Überprüfung der Reliabilität seiner Methode im Laufe der Zeit: Er verglich seine Explorationsergebnisse aus den ersten 4 Jahren (1938-1942) mit denen aus den letzten 4 Jahren (1943-1946) der Erhebungszeit, bildete unter den Gesichtspunkten des Geschlechts, der Rasse, des Familienstandes, des Al-

ters und des Bildungsgrades homogene Gruppen und verglich die Mittelwerte aller Gruppen, die mehr als 300 Personen umfaßten, aus der ersten Hälfte der Erhebungszeit mit gleichartigen Gruppen aus der zweiten Hälfte. Die Ergebnisse sind zusammengestellt in den Tabellen 21 (S. 142), 22 (S. 144) und 23 (S. 146). Die Mittelwerte aus beiden Erhebungszeiträumen sind bei allen Gruppen nahezu identisch.

„The comparisons in Tables **21-23** seem to indicate that methods of securing subjects, proficiency in interviewing, skill in using the code in which the data are recorded, and calculations and judgments which the data undergo in their statistical treatment, can be maintained at such uniform levels as many persons would have considered impossible in a case history study which is liable to error from so many sources, and which deals with as taboo a subject as sex“ (**147**).

6.2 Validität

Die Validitätsüberprüfung kann sich auf drei Stadien der Anwendung eines Erhebungsverfahrens beziehen: die Datenerhebung, die Auswertung der erhobenen Befunde (Kategorisierung, Quantifizierung, Evaluation, Scoring) und ihre Verarbeitung für eine Diagnose bzw. Prognose.

6.2.1 Die Validität der Datenerhebung

Validität der erhobenen Daten ist das Maß, in dem die von den Befragten gemachten Angaben mit einem (objektiven und seinerseits reliablen) Kriterium für den explorativ erhobenen Sachverhalt übereinstimmen.

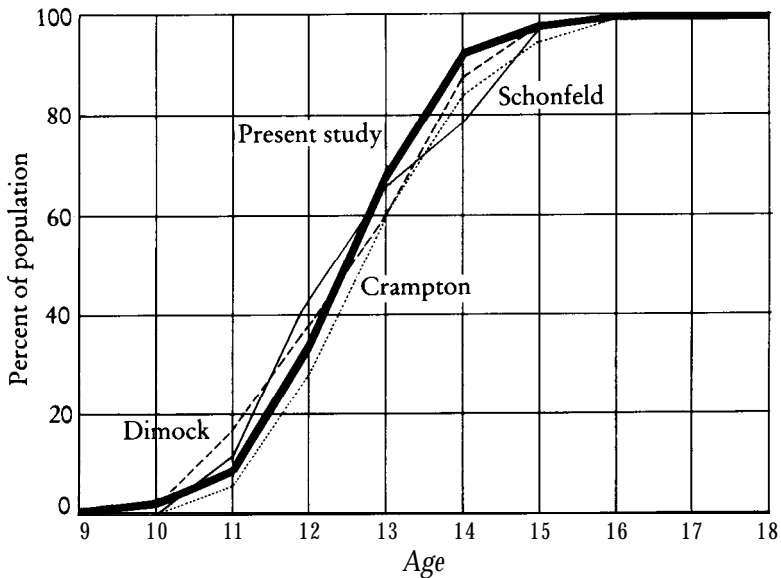
Die Leistungsfähigkeit der explorativen Verfahren bei der Datenerhebung ist selbstverständlich unterschiedlich für verschiedene Arten von Daten.

Die Leistungsfähigkeit der Exploration für die Erhebung von Daten der körperlichen Reifung in der Pubeszenz wurde von Kinsey (1948) untersucht durch einen Vergleich der explorativ erhobenen Daten mit somatoskopischen Beobachtungen. Es sei hier als Beispiel herausgegriffen das von den befragten erwachsenen Männern angegebene Alter des ersten Auftretens der Pubesbehaarung. Die sich daraus ergebende Altersverteilung wurde verglichen mit den somatoskopisch ermittelten Altersverteilungen des gleichen Merkmals, die in jenen Zeiträumen erhoben worden sind, in denen die von Kinsey befragten Männer die Pubeszenzentwicklung durchliefen. Eine frappierende Übereinstimmung ergab sich sowohl hinsichtlich der Mittelwerte (Tab. 15, S. 130) als auch bezüglich der Verteilungen (Abb. 15, S. 131).

Tabelle 1: Vergleich der Mittelwerte des ersten Auftretens von Pubesbehaarung nach den Befragungsergebnissen von Kinsey (1948) mit den Ergebnissen somatoskopischer Beobachtungen von Crampton (1908), Dimock (1937) und Schonfeld (1943)

	Crampton	Dimock	Schonfeld	Kinsey
N	3835	1406	1475	2511
M	13,44	13,08		13,45
Z			13,17	13,43
s	1.51			0,03

Auszug aus Kinsey u.a. (1948, Tab. 15, S. 130).



Wiedergabe von Kinseys (1948) Abb. 15 (S. 131).

Abb. 1: Vergleich der Altersverteilung des ersten Auftretens von Pubesbehaarung nach den Befragungsergebnissen von Kinsey (1948) mit den Ergebnissen somatoskopischer Beobachtungen von Crampton (1908), Dimock (1937) und Schonfeld (1943)

Diese gute Übereinstimmung ist deshalb so erstaunlich, weil das Alter des ersten Auftretens der Pubesbehaarung sicher zu den nicht gerade leicht erinnerbaren Ereignissen der persönlichen Entwicklung gehört.

Bezüglich des Alkoholkonsums führte Schmidt (1973) einen Vergleich von Befragungsergebnissen mit Verkaufsdaten (Kanada) durch, wobei er eine sehr gute Übereinstimmung bei niedrigen und bei mittleren Konsumquanten feststellte. Hingegen beträgt die am stärksten konsumierende Gruppe (= Personen, die 11 und mehr Flaschen alkoholischer Getränke kaufen) auf Grund der Befragungsergebnisse nur 1,8%, nach den Verkaufsdaten aber 6,54%. Besonders jene Alkoholkonsumenten, die täglich alkoholische Getränke einkaufen, machen besonders niedrige Angaben zum Konsum. Auch Boland und Roizen (1973) fanden sehr gute Übereinstimmungen zwischen Befragungsergebnissen und Verkaufsdaten bei niedrigen Konsummengen. Die Übereinstimmung wurde um so geringer, je höher die Verkaufsdaten waren. Aus diesen Ergebnissen ist zunächst einmal zu schließen, daß starke Trinker dazu tendieren, zu niedrige Konsummengen anzugeben - sei es, daß sie tatsächlich keine Kontrolle über die von ihnen genossenen hohen Alkoholmengen haben, sei es, daß sie ihren hohen Alkoholkonsum zu verschleiern trachten. Ähnliche Verhältnisse sind sicher überall zu erwarten, wo es um das Bekenntnis sozial mißbilligten Verhaltens geht. Hier zeigt sich einer der Gründe dafür, daß es gewiß auch für die explorativen Methoden eine differentielle Validität gibt: ihre Validität ist, selbst bei gleicher Thematik der Befragung, nicht für alle Menschen gleich hoch. Sie ist niedriger für sozial mißbilligte Verhaltensweisen. Dieser Tendenz gilt es, schon während der Befragung durch die bewährten Techniken der „Exploration verheimlichter Sachverhalte“ entgegenzuarbeiten. Der durch die Wirksamkeit dieser Tendenz auftretende Fehler dürfte aber für die Praxis nicht gravierend sein, denn es kommt dadurch eher ein systematischer Fehler zustande (eine speculation à la baisse bei Mengenangaben für sozial mißbilligtes Verhalten, die um so stärker ist, je höher die tatsächlichen Häufigkeiten bzw. Mengen sind), so daß die Rangfolge in den Konsumangaben und damit die hohe Reliabilität erhalten bleiben, wie Boland und Roizen (1973) für den Alkoholgenuß hervorheben.

6.2.2 Validität der Bewertung

Explorationsbefunde müssen für eine ganze Reihe von Fragestellungen in irgendeiner Form bewertet werden: sie müssen quantifiziert oder einer Auswertungskategorie zugeordnet oder irgendwie interpretiert werden.

Eine Untersuchung hierzu ist im Arbeitskreis von Thomae von Fisseni (1974) durchgeführt worden. 237 Frauen in Altersheimen wurden über das Leben im Heim befragt, und zwar einerseits in einer unstrukturierten Exploration, die einem lockeren Themenkatalog folgte, der den Tageslauf der Befragten umschrieb, und andererseits mit einem standardisierten Fragebogen, der mit 50 gezielten Fragen den gleichen Themenkreis erfaßte. Die Auswerter ordneten, unabhängig voneinander, das Explorationsmaterial dem gleichen Ja/Nein-Fra-

gebogen zu, der von den Probanden ausgefüllt worden war. Die Inter-Rater-Reliabilität betrug .73; die Übereinstimmung mit den originalen Fragebogenbeantwortungen betrug 87.

6.2.3 Validität der diagnostischen Verwertung

Die erhobenen und bewerteten Befunde bilden die Basis für die Diagnose. Die diagnostische Validität ist ein Maß für die Richtigkeit der Diagnose oder der Vorhersage. Die bewerteten Befunde sind die Prädiktoren. Voraussetzung für die Richtigkeit der Diagnosenstellung sind einerseits die Erarbeitung einer empirisch begründeten klaren Symptomatologie der zu diagnostizierenden psychologischen Sachverhalte. Eine solche ergibt sich aus empirischen Untersuchungen der Häufigkeit des Vorkommens der als Symptome aufzufassenden Befunde bei Vorliegen des zu diagnostizierenden Sachverhaltes. In mustergültiger Weise hat Kunkel (1977) dieses Vorgehen für die Prognose der Rückfallwahrscheinlichkeit in das Delikt des Führens eines Kraftfahrzeuges in alkoholbedingt fahruntüchtigem Zustand vorgeführt. Man kann diesen Teil der Validität als „symptomatische Bedeutsamkeit“ (Wohlfahrt, 1939) oder als „diagnostische Validität“ bezeichnen. Sodann ist (weitere) Voraussetzung der richtigen Diagnosenstellung, daß der Diagnostiker die „symptomatische Bedeutsamkeit“ der erhobenen Befunde kennt, d.h. sie richtig gewichtet und sie im „semantischen Umfeld“ (K. Bühler, 1933) richtig interpretiert.

In einem unveröffentlichten Versuch von U. Lehr (Thomae 1968, 117f.) waren 46 Studenten der Psychologie Ausschnitte aus der Exploration eines 38jährigen Kaufmanns, dem wegen Betrugs und Unterschlagung die FE entzogen worden war, vom Tonband dargeboten worden, und zwar a) über Kindheit und Jugend des Probanden und b) zu den Straftaten und seinen Auseinandersetzungen mit dem Straßenverkehrsamt. Die Beurteiler hatten die Aufgabe, den Eindruck wiederzugeben, den sie daraus von der Persönlichkeit des Probanden erhielten. Die von mindestens der Hälfte der Beurteiler übereinstimmend gegebenen Charakterisierungen wurden verglichen mit den Aussagen über diese Persönlichkeit, wie sie auf Grund der psychologischen Untersuchung, bei der u.a. angewendet worden waren: Kieler Determinationsgerät, Hawie, Rorschach, TAT, graphologische Analyse, gewonnen wurden. Der Vergleich der Aussagen, die die Beurteiler auf Grund der Exploration über Kindheit und Jugend überwiegend übereinstimmend machten, mit den Befunden aus der persönlichkeitsdiagnostischen Untersuchung ergab bei 24 Aussagen völlige Übereinstimmung, bei 16 Aussagen war eine Bestätigung aus den anderen diagnostischen Verfahren nur teilweise zu erhalten, bei 2 Aussagen widersprachen sich der Eindruck aus der Exploration und die anderen diagnostischen Befunde. Im Vergleich der Aussagen, die die Beurteiler auf Grund der Exploration zur Sache (= seine bisherige Verkehrsbewährung) machten, mit den

Befunden aus der Persönlichkeitsuntersuchung ergaben sich bei 18 Aussagen Übereinstimmungen, für 19 Aussagen waren entsprechende Befunde aus den anderen diagnostischen Verfahren nicht beizubringen, bei 6 Aussagen bestand ein Widerspruch zu den Befunden aus den anderen diagnostischen Verfahren. Thomae 1968 interpretiert das Ergebnis dieses Versuches wie folgt:

„Die in der Exploration gegebene Selbstdarstellung des Probanden ist nicht nur das Resultat von Rationalisierungen, von Vereinfachungen und Verfälschungen. Insbesondere dort, wo konkrete Verhaltensweisen in konkreten Situationen geschildert werden, wird eine psychische Realität erkennbar, die andere Verfahren niemals in dieser Eindeutigkeit aufzuweisen vermögen . . . Soweit aus solchen Schilderungen Verhaltenstendenzen erschlossen werden, lassen sich diese wenigstens zum erheblichen Teil auch durch ‚objektivere‘ Verfahren nachweisen. Bei anderen dürfte die Exploration auf Tendenzen verweisen, die den übrigen Verfahren gar nicht zugänglich werden.

Darüber hinaus zeigt das Ergebnis, daß das Maß der Übereinstimmung zwischen der Verwertung der Explorationsdaten durch Beurteiler und den Testbefunden von dem jeweiligen Inhalt der Mitteilung abhängt“ (118).

Die von Thomae entwickelten Prinzipien einer Analyse der formalen Qualitäten des Tageslaufes wurden in einer (unveröffentlichten) Untersuchung von Olbrich (1967) auf ihren Aussagewert hin überprüft (Thomae 1968). Das Material bildeten 40 in der Exploration gegebene Schilderungen von Tagesläufen an einem „typischen Werktag“ von 60- bis 75jährigen Männern und Frauen (erhoben im Rahmen der VW-Alters-Studie). Zwei Beurteiler stuften, unabhängig voneinander, die Schilderung des Verhaltens der befragten Personen in 8 formalen Kategorien des Verhaltens (Thomae 1968, Tab. 5, S. 130) auf einer 9stufigen Skala ein, ohne sonst irgendeine Kenntnis von den untersuchten Personen zu haben. Zwischen den Werten der 8 hier verwendeten Ausgangsskalen und den übrigen in dieser Untersuchung erhobenen Daten (Gesamtzahl 563) über die Erzähler dieser Tagesläufe ergaben sich bei Anwendung der Rangreihenkorrelation mehr als 300 signifikante bis sehr signifikante Korrelationen. Die Ergebnisse von Olbrich belegen, daß eine methodisch fundierte Erhebung und Auswertung von Daten über Verhaltensweisen in mittleren biographischen Einheiten (Tageslauf) symptomatische Bedeutsamkeit besitzt sowohl für Varianten des Verhaltens in kurz-zeitigen Einheiten (z.B. Leistung beim Determinationsgerät nach Mierke) wie bei größeren Einheiten der Biographie (z.B. Formen der Auseinandersetzung mit einer beruflichen oder familiären Situation).

Die Bedeutung der Exploration für die Eignungsdiagnostik ist sehr schwer konkret nachweisbar, weil in der Praxis in der Regel das Eignungsurteil nicht allein auf Grund von Explorationsbefunden abgegeben wird, sondern auch die äußere Erscheinung des Bewerbers, die Verhaltensbeobachtung, Informationen aus Zeugnissen und Akten, meistens auch Testergebnisse in das abschließende Eignungsurteil eingehen. Bei Bewährungskontrollen der in der Ober-

gutachterstelle zur Beurteilung der Kraftfahreignung des Landes Nordrhein-Westfalen (Leiter: Undeutsch) verwendeten Untersuchungsverfahren war es möglich, einige ausschließlich aus der Exploration gewonnene Daten auf ihren prognostischen Wert hin zu untersuchen. Welzel (1976 und 1982) verfolgte an einer Untersuchungsgruppe von 420 (meist mehrfach) wegen Trunkenheit am Steuer vorbestraften Probanden, die in der Obergutachterstelle untersucht worden waren, deren Verkehrsbewährung in den auf die Untersuchung folgenden Jahren. Von den ausschließlich explorativ erhobenen Daten stehen das „Alter zum Zeitpunkt der ersten Heirat“ und der „größte Alkoholkonsum bei einem Trinkanlaß im letzten Jahr vor der Untersuchung“ in gesicherter Beziehung zu einem erneuten Rückfall in ein Trunkenheitsdelikt beim Führen von Kraftfahrzeugen innerhalb von 3 Jahren nach Wiedererteilung der Fahrerlaubnis ($\Phi = .36$). Lediglich einige aus der Gesamtheit der Untersuchungsergebnisse abgeleitete Persönlichkeitseigenschaften und durch Aktenanalyse erhobenen Prädiktoren erreichen höhere oder ähnlich hohe Validitätskoeffizienten.

Nicht mehr exakt faßbar, aber doch immerhin deutlich erkennbar ist die Rolle der Exploration in der Eignungsdiagnostik hervorgetreten in einer Bewährungsuntersuchung von Trankell (1956). Das Material der Bewährungskontrolle waren Bewerber um Einstellung als Flugzeugführer bei der schwedischen Luftwaffe. Die Eignung wurde von den Psychologen auf einer 9stufigen Skala angegeben, wobei sowohl die Testergebnisse als auch der in der Exploration gewonnene Eindruck berücksichtigt wurden. Als Kriterium diente die

Tabelle 2: Korrelationen der Untersuchungsvariablen mit der Endbeurteilung nach Abschluß der fliegerischen Grundausbildung

Variable	1945	1946	1947	1948
Techn. Verständnis		.42	.28	.28
Flugtechn. Orientierung			.25	.05
Allgemeine Flugorientierung		.24	.34	.05
Sterzinger, Qualität,	.09	.18	.27	.09
Sterzinger, Quantität	.00	.22	.31	.17
Formale Intelligenz	.25	.31	.47	.27
Introversion	.16	.01	.06	.21
Sensibilität		.04	.16	.13
Körperliche Gewandtheit			.30	.10
Eignungsurteil	.39	.08	.57	.56
N		96	96	100

(Aus: Trankell 1956, Tab. 17, S. 84)

Bewertung, die die angenommenen Kandidaten am Ende der fliegerischen Grundausbildung erhielten. Die Ergebnisse sind in Tab. 2 wiedergegeben: Korrelationen der Untersuchungsvariablen mit der Endbeurteilung nach Abschluß der fliegerischen Grundausbildung.

Es zeigt sich, daß die von den Psychologen vorgenommenen Einstufungen mit dem Kriterium - mit Ausnahme des Aufnahmejahrganges 1946 - höher korrelieren als die Testwerte. Die aus der Exploration gewonnenen Zusatzinformationen und der auf Grund der Exploration gebildete Gesamteindruck ermöglichen eine genauere Prognose als irgendeiner der Tests. Eine Ausnahme bildet der Aufnahmejahrgang 1946. Eine Erklärung dafür findet sich, wenn man die Validitätskoeffizienten für die beteiligten Psychologen einzeln berechnet (Tab. 3).

Tabelle 3 : Korrelationen der Eignungsurteile der einzelnen Psychologen mit der Endbeurteilung nach Abschluß der fliegerischen Grundausbildung

	Explorator	N	Eignungsurteil
1946	{ Psychologe A	69	.21
	{ Psychologe B	27	.00
1947	{ Psychologe C	69	.33
	{ Psychologe D	27	.50
1948	{ Psychologe E	39	.36
	{ Psychologe F	61	.49

(Aus: Trankell 1956, Tab. 19, S. 87)

Diese Zusammenstellung offenbart beträchtliche Unterschiede in der prognostizierenden Befähigung der beteiligten Psychologen. Die Eignungsprognosen des Psychologen B haben keinerlei prognostischen Wert, während die Psychologen F und D sehr hohe Validitätskoeffizienten erzielten. Die Leistungsunterschiede zwischen den beteiligten Exploratoren traten noch deutlicher in Erscheinung, wenn als Validitätskriterium das Alternativmerkmal „fliegerische Grundausbildung wegen mangelhaften Ausbildungserfolges abgebrochen ./ erfolgreich abgeschlossen“ verwendet wird. Die individuellen Validitätskoeffizienten der beteiligten Psychologen lagen hier zwischen .40 und der (nicht unbeträchtlichen) negativen Korrelation von -.21 (Tab. 18, S. 87). Die Auslese im Jahre 1946 wurde von den zwei Psychologen A und B vorgenommen, die dazu am wenigsten befähigt waren.

Ein ähnliches Ergebnis erbrachte die Bewährungskontrolle der von der Obergutachterstelle des Landes Nordrhein-Westfalen zur Beurteilung der Kraftfahreignung erstellten psychologischen Gutachten (Deters, 1978). 878 Probanden, die in den Jahren 1966-1972 begutachtet worden waren, wurden über einen Zeitraum von 3,5-7,0 Jahren bezüglich ihrer späteren Verkehrsbewährung weiter verfolgt. In den Untersuchungen der Obergutachterstelle nimmt die Exploration eine zentrale Stellung ein. Sie erstreckt sich sowohl auf die Persönlichkeit des Probanden wie auch auf seine bisherige Verkehrsbewährung und bei Trunkenheitstätern selbstverständlich auch auf die Entwicklung ihrer Trinkgewohnheiten und ihrer Einstellungen zum Problemkreis „Alkohol und Verkehr“. Die durchschnittliche Dauer einer Exploration beträgt hier 2;15 Stunden. Bei der Prognose künftiger Verkehrsbewährung haben die aus der Exploration gewonnenen Befunde ein sehr starkes Gewicht. Es wurde nach verschiedenen statistischen Verfahren die prospektive Validität der in die Begutachtung eingegangenen Variablen ermittelt. Es wurden auch multiple Konstriktionskoeffizienten berechnet, um die optimale Kombination von Variablen für die Prognosestellung kennenzulernen. U. a. wurde eine Trennung der Ergebnisse nach Kombinationen von Prädiktorvariablen mit und ohne Einbeziehung der Variablen „Gutachtenergebnis“ vorgenommen. Es ergab sich, daß die Sechser-Kombinationen, bei denen das Gutachtenergebnis einbezogen worden war,

„fast durchweg numerisch höhere Vorhersageverbesserungen erzielen als die Sechser-Kombinationen, bei denen die Variable ‚Gutachtenergebnis‘ ausgespart worden war“ (114).

Es ergab sich weiter, daß bei allen als optimal gefundenen Prädiktorkombinationen der jeweils beste Einzelprediktor die Variable „Gutachtenergebnis“ ist (S. 16 und Tab. 6/B 1).

Dieses immer wieder zu beobachtende diagnostische Plus, das die Exploration liefert, hat zwei Gründe. Einerseits erfaßt die Exploration inhaltliche Aspekte der Persönlichkeit, die andere Verfahren niemals in dieser Eindeutigkeit aufzuweisen vermögen. Zum anderen ist sie ein ganzheitliches Verfahren, das es ermöglicht, die in der Exploration und die mit anderweitigen Untersuchungsverfahren gewonnenen Befunde aufeinander zu beziehen und in einer im Hinblick auf die konkrete Persönlichkeit adäquaten Weise zu gewichten. Wellek (1958) hatte schon auf Grund theoretischer Überlegungen darauf hingewiesen, daß die Exploration, je mehr sie „die Persönlichkeit als Ganzes im Auge hat“,

„desto mehr leistet die Methode und desto eher wird sie zum ‚Rückgrat‘ der Diagnose, die dann eben das Ganze, nicht ausgegliederte Sektoren betrifft“ (24).

Noch lebendiger und anschaulicher schildert den unersetzlichen Wert der Exploration Walther (1941) aus den Erfahrungen der ganzheitlich-charakterolo-

gisch ausgerichteten Offiziersbewerber-Eignungsuntersuchungen in der deutschen Wehrmachtpsychologie:

„Die Eignung der Aussprache zur Lösung der verschiedenartigsten diagnostischen Probleme ist nicht von ungefähr. Es würde eine Abhandlung für sich erfordern, das im einzelnen darzulegen. In diesem Zusammenhang mag es genügen, daran zu erinnern, daß wir in der Aussprache die einzigartige Möglichkeit besitzen, jede Reaktion des Prüflings in ihrer Bedeutung für das Ganze der Persönlichkeit genau abzuschätzen. Seine Äußerungen sind nicht (wie etwa schriftliche) von ihm abgelöst, sondern bleiben im Verband nicht nur der zugehörigen Ausdruckerscheinungen und Verhaltensweisen, sondern auch der situativen Faktoren. Schließlich lassen sich die tieferen inneren Zusammenhänge, in denen die Reaktionen und Äußerungen des Prüflings stehen, grundsätzlich bis an die Grenze des Möglichen erkunden. Aus diesem Grunde ist die Exploration die *via regia* der psychologischen Diagnostik und wird es immer bleiben. Das heißt mit anderen Worten auch: es wird gar nicht möglich sein, eine ‚bessere‘ Methode zu finden“ (24).

Auf der anderen Seite gibt es auch niederschmetternde Validitätskoeffizienten aus dem Bereich der Eignungsdiagnostik. Solche finden sich bei Eysenck (1951, 1952) zusammengestellt, aber auch in allen Sammelreferaten über die Leistungsfähigkeit explorativer Techniken in der Eignungsdiagnostik (Wagner 1949, Mayfield 1964, Ulrich & Trumbo 1965, Moffatt 1969, Wright 1969, Triebe 1976). Wegen der vielfältigen Verwendungsmöglichkeit explorativer Techniken ist es ganz selbstverständlich, daß es auch eine Vielzahl, und zwar sehr unterschiedlicher Validitätskoeffizienten geben muß. Niedrige Koeffizienten besagen zunächst einmal, daß Explorationen - im Gegensatz zu voll standardisierten Tests - auch ganz unzulänglich durchgeführt und ausgewertet werden können. Wellek (1958) sagt zu den von Eysenck berichteten niedrigen Validitätskoeffizienten:

„Wie schlecht muß da exploriert worden sein, wenn keinerlei Bewährungserfolg erzielt werden konnte!“ (26).

Auch Thomae (1968) meint, Eysencks

„diesbezüglichen Befunde sprechen zunächst einmal für einen nicht sonderlich hohen Standard der Verwendung von Explorationstechnik und der Techniken und Ziele der Auswertung der auf diese Weise gewonnenen Daten“ (117).

Zum anderen stellt sich das - für alle Validitätsuntersuchungen leidige - Problem der Kriterienvariablen im Falle eines ganzheitlichen Verfahrens, wie es die Exploration in typischer Weise ist, mit besonderer Schärfe. Hierzu hatte Wellek schon 1958 ausgeführt:

„Eine Exploration kann selbst im günstigsten Falle das Ergebnis eines Tests nicht besser voraussagen als dieser selbst. Ist also der der Bewährungskontrolle zugrunde gelegte Maßstab ein Test oder doch testartig, dann ist es ein methodischer Kurzschluß, zu sagen, in der Bewährung komme ein Test besser heraus als eine Exploration, denn das

liegt in der Sache selbst. Ein ‚ganzheitliches‘ Verfahren wie die (echte) Exploration kann der Natur der Sache nach nur überlegen, ja sogar nur brauchbar sein, wenn das Erkenntnisziel - der gefragte Gegenstand - gleichfalls ein ganzheitliches ist“ (27).

„Allgemeiner gesagt: eine nicht qualitativ ausgerichtete Bewährungskontrolle ist zur Validierung eines ganzheitlichen diagnostischen Verfahrens nicht geeignet“ (28).

Wahrscheinlich in noch viel höherem Maße ist ein anderer Grund für die unterschiedliche Höhe der gefundenen Validitätskoeffizienten maßgebend: Lopez (1965) weist darauf hin, daß die Eignungsdiagnose ein mehrstufiger Vorgang ist:

„This decision-making function of the employment interviewer presupposes three separate and prior steps: description, evaluation, and prediction. The selection interviewer must first elicit sufficient information from the applicant (description) to compare with a set of preestablished job specifications (evaluation) to enable him to draw a conclusion about the probable future behavior of the interviewee in a specific set of circumstances (prediction). On the basis of this prediction he then makes a decision ...“ (10f.).

Für die diagnostische Valenz der Exploration ist von ausschlaggebender Bedeutung, daß für die Diagnose eine klare und vollständige Symptomatik der zu diagnostizierenden Sachverhalte von der Forschung erarbeitet worden ist, der nicht nur zu entnehmen ist, welche Befunde überhaupt symptomatische Bedeutsamkeit besitzen, sondern vor allem auch, welches Gewicht den einzelnen Symptomen beizumessen ist. Das Fehlen einer solchen klaren Symptomatik ist z.B. der Grund dafür, weshalb selbst bei standardisierten Tests und Persönlichkeitsfragebögen in der Klinischen Psychologie von verschiedenen Auswertern unterschiedliche klinische Diagnosen gestellt werden. Für die Eignungsdiagnostik haben Triebe, Fischer und Ulich (1973) mit Recht hervorgehoben: Eine Analyse der Arbeitsanforderungen

„ist der Ausgangspunkt jeder den Ansprüchen wissenschaftlicher Objektivität genügenden Eignungsdiagnostik. Ihr kommt die Schlüsselfunktion zu; von ihrer Angemessenheit und Genauigkeit hängt letzten Ende ab, ob sich die eignungsdiagnostischen Verfahren - mit den resultierenden Prognosen - bewähren“ (27).

Niedrige Validitätskoeffizienten können ihren Grund ganz einfach haben in unzutreffenden Vorstellungen von der Anforderungsstruktur der betreffenden Stelle oder Position auf Seiten des Entscheidungsträgers. Die Strukturierung der eignungsdiagnostischen Exploration muß von einer Analyse der Arbeitsanforderungen her vorgenommen werden, während in der Praxis, wie die vorliegende Literatur zeigt, weit eher ein dem Explorator mehr oder weniger bewußtes „Stereotyp des guten Bewerbers“ den Orientierungsrahmen für die Gesprächsführung und für die Beurteilung abgibt.

„Insofern könnte man - etwas überspitzt - sagen, daß mit Hilfe unstrukturierter Interviews ausgewählte Bewerber zwar wahrscheinlich meist die Anforderungen ihres

Interviewers erfüllen werden, daß es aber weitgehend vom Zufall abhängen wird, in welchem Maße dessen Anforderungen mit denen der für den Bewerber vorgesehenen Tätigkeit übereinstimmen“ (Triebe 1976, 40).

Aus den Sammelreferaten von Ulrich und Trumbo (1965) und von Wright (1969) ergibt sich eine Überlegenheit strukturierter Explorationen hinsichtlich ihrer Reliabilität und Validität. Dies gilt aber wiederum nur, wenn sich die Voraussage auf etwas Spezialisiertes, Eingegengtes bezieht. Allgemein formuliert Wellek (1958):

„Je enger die Fragestellung, zumal in Richtung auf konkrete Leistungen und Erfolge, umschrieben ist, je weniger läßt sich von der bloßen Exploration, je mehr von entsprechend zugeschnittenen Tests und sogar Fragebögen erwarten. Und umgekehrt: je allgemeiner aufs Persönlichkeitsganze gehend die Fragestellung, je mehr läßt sich von der Exploration und von der unmittelbaren oder auch mittelbaren ausdrucksmäßigen Kenntnisnahme, je weniger von Tests und erst recht von Fragebögen erwarten“ (25).

Auf Grund einer Übersicht über die vorliegende anglo-amerikanische Literatur zur Leistungsfähigkeit der Exploration im Kontext der Eignungsdiagnostik kommt Triebe (1976) zu dem Ergebnis:

„Im Hinblick auf die Möglichkeiten des Interviews scheint-wenn auch wohl z.T. aus recht verschiedenen Gründen - unter Praktikern und Wissenschaftlern fast einheitlicher Optimismus zu herrschen“ (8).

Diese Möglichkeiten gilt es zu nutzen. Für die vielfältigen Anwendungsbereiche explorativer Techniken gilt es,

die relevanten Themen der Exploration empirisch zu ermitteln,
Fragetechniken zu entwickeln,
Auswertungstechniken zu erarbeiten,
die diagnostische Relevanz von Explorationsdaten zu ermitteln,
Psychologen auf Grund dieser Erkenntnisse in Explorationstechnik und
-auswertung auszubilden,

damit der Kreis von Psychologen, der mit diesem methodischen Instrumentarium erfolgreich zu arbeiten versteht, vergrößert wird, denn dieses Erhebungsverfahren hat einzigartige und daher unverzichtbare Vorteile, auf die namentlich Thomae immer wieder hingewiesen hat:

„Es ist somit nicht eine methodische Voreingenommenheit oder Borniertheit, die uns die besondere Bedeutung der Exploration für die systematische Beobachtung menschlichen Verhaltens in verschiedenen biographischen Einheiten hervorheben läßt. Vielmehr ist es die Einsicht, daß nur das Individuum selbst Zeuge seines Verhaltens im natürlichen Ablauf seines Lebens ist. Da wir keine Zeiten herbeisehnen dürften, in denen eine Dauerbeobachtung durch Fremde staatlich oder wissenschaftlich sanktioniert wird, können wir auf die Aussagen dieses Zeugen nicht verzichten“ (1968, 111).

Literatur

- Anger, H. 1969. Befragung und Erhebung. Handbuch der Psychologie, Band 7/1: Sozialpsychologie, Göttingen: Hogrefe, 567-617.
- Antons, K. & Schulz, W. 1976. Normales Trinken und Suchtentwicklung. Band 1, Göttingen: Hogrefe.
- Arnold, W. & Pauli, R. 1957⁶. Psychologisches Praktikum. Band 1, Stuttgart: Fischer.
- Assessment Staff 1948. Assessment of men: selection of personnel for the office of strategic services. New York: Rinehart.
- Baade, W., Lipmann, O. & Stern, W. 1909. Fragment eines biographischen Themas. Zeitschrift für Angewandte Psychologie, 3, 191-215.
- Bain, R. K. 1960. The researcher's role: a case study. In: Adams, R. N. and Preuss, H. H. (eds): Human Organisation research. Field relations and techniques, Homewood: Dorsey Press, 140-152.
- Beck, W. 1942. Begegnung und Erkenntnis. Zeitschrift für Angewandte Psychologie, 62, 328-369.
- Bellows, R. 1954. Psychology of Personnel in Business and Industry. Englewood Cliffs: Prentice-Hall.
- Bellows, R. M. & Estep, M. F. 1954. Employment psychology: The interview. New York: Holt, Rinehart & Winston.
- Bingham, W. van Dyke & Moore, B. V., with the collaboration of Gustad, J. W., 1959. How to interview. New York: Harper & Brothers.
- Blakeney, R. N. & McNaughton, J. F. 1971. Effects of temporal placement of unfavorable information on decision making during the selection interview. Journal of applied Psychology, 55, 138-142.
- Böcher, W. 1968. Befragungsstil und Verdeckungsmöglichkeiten im eignungsdiagnostischen Gespräch. Fahreignung und Verkehrssicherheit, Mitteilungsblatt des MPI Stuttgart, 17, 20-25.
- Boland & Roizen 1973 zit. nach Antons u. Schulz 1976.
- Bolton, D. L. & Hickey, M. E. 1969. Effect of interviews on teacher selection decisions. Journal of Applied Psychology, 53, 501-505.
- Bühler, K. 1933. Ausdruckstheorie. Jena: Fischer.
- Cattell, R. B. 1957. Personality and motivation: structure and measurement. Yonkers-on-Hudson, New York: World Book.
- Cosper, R. 1969. Interviewer bias in a study of drinking practices. Quarterly Journal of Studies on Alcohol, 30, 152-157.
- Crampton, C. W. 1908. Psychological age a fundamental principle. Child Development, 15, 3-52.
- Deters, H. 1978. Beiträge zur Prognose der Eignung zum Führen von Kraftfahrzeugen. Unveröff. vervielf. Forschungsbericht der Arbeits- und Forschungsgemeinschaft für Straßenverkehr und Verkehrssicherheit. Köln.

- Dimock, H. S. 1937. Rediscovering the adolescent. A study of personality development in adolescent boys. New York: Association Press.
- Ekman, G. 1947. Reliabilitet och konstans. Uppsala: Hugo Gebers Förlag.
- Erbslöh, E. 1972. Interview. Stuttgart: Teubner Studienskripten.
- Eysenck, H.-J. 1951. Uses and abuses of psychology. London: Penguin Books; deutsche Übersetzung: Wege und Abwege der Psychologie. Hamburg 1956.
- Fear, R. A. 1953. The evaluation interview. New York: McGraw Hill.
- Fisseni, H.-J. 1974. Zur Zuverlässigkeit von Interviews. Archiv für Psychologie, 126, 71-84.
- Fleiss, J. L. 1970. Estimating the reliability of interview data. Psychometrika, 35, 143-162.
- Friedrichs, H. 1967. Die aussagepsychologische Exploration. Handbuch der Psychologie, Band 11, Göttingen: Hogrefe, 3-25.
- Fröhlich, C. P. 1958. The completeness and accuracy of counseling interview reports. Journal of General Psychology, 58, 81-96.
- Fröhlich, R. E. & Bishop, F. M. 1973. Die Gesprächsführung des Arztes. Berlin: Springer.
- Gross, H. 1893. Handbuch für Untersuchungsrichter, Polizeibeamte und Gendarmen usw. als System der Kriminalistik. Graz: Leuschner & Lubensky.
- Gross, H. 1898. Kriminal-Psychologie. Leipzig: Vogel.
- Husén, T. 1949. Om innebörden av psykologiska mätningar. Lund-Köbenhavn: Gleerup och Munksgaard.
- Inbau, F. E. & Reid, J. E. 1967², 1974 reprint. Criminal interrogation and confessions. Baltimore: The Williams & Wilkins Company.
- Jung, C. G. 1937. Die Beziehungen der Psychotherapie zur Seelsorge. Zürich: Rascher. Wiederabgedruckt in: Gesammelte Werke, 11. Bd., 2. Aufl. 1979, S. 355-376.
- Kahn, R. L. & Cannell, C. F. 1966¹⁰. The dynamics of interviewing. New York: John Wiley & Sons.
- Kemmler, L. & Echelmeyer, L. 1978. Anamnese-Erhebung. Handbuch der Psychologie, Band 8/2, Göttingen: Hogrefe, 1628-1648.
- Kephart, N. 1952. The employment interview in industry. New York: McGraw-Hill.
- Kinsey, A. C., Pomeroy, W. B. & Martin, C. E. 1948. Sexual behavior of the human male. Philadelphia and London: Saunders; deutsche Übersetzung: Das sexuelle Verhalten des Mannes. Berlin und Frankfurt 1964.
- Kinsey, A. C., Pomeroy, W. B., Martin, C. E. & Gebhard, P. H. 1953. Sexual behavior in the human female. Philadelphia and London: Saunders; deutsche Übersetzung: Das sexuelle Verhalten der Frau. Berlin und Frankfurt 1963.
- Klages, L. 1926. Zur Ausdruckslehre und Charakterkunde. Heidelberg: Kampmann.
- Klebsberg, D. v. 1970. Fahrverhalten. Kleine Fachbuchreihe, Band 8, Wien: Kuratorium für Verkehrssicherheit.

- Kreipe, K. 1936. Zur Methode der Exploration. In: Abhandlungen zur Wehrpsychologie, 1. Folge, Zeitschrift für Angewandte Psychologie und Charakterkunde, Beiheft 72, 104-114.
- Kröber, W. 1942. über die Hauptaussprache (Exploration). Wehrpsychologische Mitteilungen, 4/1, 26-39.
- Kunkel, E. 1977. Biographische Daten und Rückfallprognose bei Trunkenheitstätern im Straßenverkehr. Köln: Verlag TÜV Rheinland.
- Kunkel, E. 1978. Die Bedeutung der biographischen Daten für die Fahreignungsprognose. Köln: Verlag TÜV Rheinland.
- Ledvinka, J. 1971. Rate of interviewer and the language elaboration of black interviewees. Journal of Social Issues, 27, 185-197.
- Ledvinka, J. 1972. Rate of employment interviewer and reasons given by black job-seekers for leaving their jobs. Proceedings of the Annual Convention of the American Psychological Association, 7, 441-442.
- Lehr, U. 1964. Diagnostische Erfahrungen aus explorativen Untersuchungen bei Erwachsenen. Psychologische Rundschau, 14, 97-106.
- Lewin, K. 1936. Principles of Topological Psychology, New York-London: McGraw-Hill. Deutsche Übersetzung: Grundzüge der topologischen Psychologie. Bern-Stuttgart-Wien: Huber 1969.
- Lienert, G. A. 1969³. Testaufbau und Testanalyse. Weinheim: Beltz.
- Lopez, F. M. 1965. Personnel interviewing. New York: McGraw-Hill.
- Magnusson, D. 1959. A study of ratings based on T. A. T. Stockholm: The Swedish Council for Personnel Administration.
- Margis, P. 1911. Das Problem und die Methoden der Psychographie. Zeitschrift für Angewandte Psychologie, 5, 409-451.
- Mayfield, E. C. 1964. The selection interview: A re-evaluation of published research. Personnel Psychology, 17, 239-260.
- McFarlane, J. W. 1938. Studies in child guidance. I. Methodology or data collection and Organisation. Society for Research in Child Development, 3, 1-254.
- Meier, R. D. 1972. The effectiveness of modeling procedures and instruction for teaching verbal employment interview behaviors to high school seniors. Dissertation Abstracts. 33, (6-B), 2817.
- Metzger, W. 1942. Psychologie und Menschenkenntnis. Die Erziehung. Leipzig: Quelle & Meyer.
- Michel, L. 1969. Empirische Untersuchungen zur Frage der Übereinstimmung und Gültigkeiten von Beurteilungen des intellektuellen Niveaus aus der Handschrift. Archiv für die gesamte Psychologie, 121, 31-54.
- Mierke, K. 1944. Psychologische Diagnostik. In: N. Ach (Hrsg.): Lehrbuch der Psychologie, 3. Bd., Bamberg: Buchner, 1-79.

- Moffatt, G. W. 1969. The selection interview. A review. *Personnel Practice Bulletin*, **25**, 15-23.
- Olbrich, M. 1967. Formale Analyse der Verhaltensstruktur auf Grund des Tageslaufes. Vordiplomarbeit Bonn.
- Oldfield, R. C. 1951⁴. The psychology of the interview. London: Methuen.
- Orne, M. T., Thackray, R. I. & Paskewitz, D. A. 1972. On the detection of deception. In: N. S. Greenfield and R. A. Sternbach (eds): *Handbook of Psychophysiology*. New York: Holt, Rinehart and Winston, 743-785.
- Parkinson, C. N. 1957. *Parkinsons Law*. Boston: Houghton Mifflin Comp. Deutsche Übersetzung: *Parkinsons Gesetz*. Stuttgart 1958.
- Pfahler, G. 1939. Das Gespräch als Methode erbcharakterologischer Rassenforschung. Bericht über den XVI. Kongreß der Deutschen Gesellschaft für Psychologie, **119-122**.
- Pongratz, L. 1957. Das psychologische Explorationsgespräch. *Psychologische Rundschau*, 8, 195-205.
- Rechetnick, J. & Barkus, Ph. 1966. The anatomy of a Workshop. *Personnel Review*, **27**, 199-205.
- Richardson, S. A., Dohrenwend, B. S. & Klein, D. 1965. *Interviewing*. New York: Basic-Books.
- Roethlisberger, F. J. & Dickson, W. J. 1939. *Management and the worker*. New York: Wiley.
- Rowe, P. M. 1963. Individual differences in selection decisions. *Journal of Applied Psychology*, 47, 304-307.
- Salber, W. 1960. Qualitative Methoden der Persönlichkeitsforschung. In: *Handbuch der Psychologie*, Band 4, Göttingen: Hogrefe, 30-58.
- Salber, W. 1969. Strukturen der Verhaltens- und Erlebensbeschreibung. In: *Enzyklopädie der Geisteswissenschaftlichen Arbeitsmethoden*, München und Wien: Oldenbourg, 52.
- Scheuch, E. 1962. Das Interview in der Sozialforschung. In: R. König (Hrsg.): *Handbuch der empirischen Sozialforschung*, Stuttgart: Enke.
- Schmidt 1973. zit. nach Antons u. Schulz 1976.
- Schonfeld, W. A. 1943. Primary and secondary sexual characteristics. Study of their development in males from birth through maturity with biometric study of penis and testes. *American Journal for the Diseases of the Child*, 65, 535-549.
- Schraml, W. 1964. Das psychodiagnostische Gespräch (Exploration und Anamnese). *Handbuch der Psychologie*, Band 6, Göttingen: Hogrefe, 868-897.
- Schuller, A. & Rosenmeier, H. P. 1973. *Medizinstudium und Sozialwissenschaften*. München: Urban & Schwarzenberg.
- Soudijn, K. A., Mellenbergh, G. J. & Hartemink, B. G. 1970. Evaluatie van een handleiding voor de interviewer. *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden*, 25, 618-626.

- Spiegel, B. 1965. Die Aufgaben der Psychologie in der werbewissenschaftlichen Forschung. *Werbewissenschaftliches Referatenblatt*, Nr. 3.
- Spriegel, W. R. & James, V. A. 1958. Trends in recruitment and selection practices. *Personnel*, 35, 42-48.
- Steinkamp, S. W. 1966. Some characteristics of effective interviewers, *Journal of Applied Psychology*, 50, 487-492.
- Stern, W. 1900. Die differentielle Psychologie in ihren methodischen Grundlagen. Leipzig: Barth.
- Stern, W. 1902. Zur Psychologie der Aussage. *Zeitschrift für Strafrechtswissenschaft*, **22**, 315-370.
- Taft, R. 1955. The ability to judge people. *Psychological Bulletin*, 52, 1-23.
- The McGraw-Hill course in effective interviewing. New York: McGraw-Hill 1973.
- Thomae, H. 1968. Das Individuum und seine Welt. Göttingen: Hogrefe.
- Thomae, H. & Schmidt, H. D. 1967. Psychologische Aspekte der Schuldfähigkeit. *Handbuch der Psychologie*, Band 11, Göttingen: Hogrefe, 326-396.
- Thoms, K. 1975. Anamnese. In: S. Keil (Hrsg.): Familien- und Lebensberatung. Ein Handbuch. Stuttgart: Huber.
- Thorne, F. C. 1970. Psychological „twenty questions“: A method for teaching diagnostic interviewing. *Journal of Clinical Psychology*, 26, 331-334.
- Trankell, A. 1956. Rekryteringen av piloter i svenska flygvapnet. *Tidskrift i militär hälsovård*, Jg. 81, 61-90.
- Trankell, A. 1959. The psychologist as an instrument of prediction. *Journal of Applied Psychology*, 43, 170-175.
- Triebe, J. K. 1976. Das Interview im Kontext der Eignungsdiagnostik. Bern, Stuttgart, Wien: Huber.
- Triebe, J. K., Fischer, K. & Ulich, E. 1973. Problemstudie zur Informations- und Entscheidungsfindung bei der Auswahl von Bewerbern für den öffentlichen Dienst. In: Studienkommission für die Reform des öffentlichen Dienstrechts, Band 10, 15-104, Baden-Baden.
- Ulrich, L. & Trumbo, D. 1965. The selection interview since 1949. *Psychological Bulletin*, 63, 100-116.
- Undeutsch, U. 1954. Die Entwicklung der gerichtropsychologischen Gutachtertätigkeit. Bericht über den XIX. Kongreß der Deutschen Gesellschaft für Psychologie, Göttingen, 132-154.
- Undeutsch, U. 1965². Forensische Psychologie. In: *Handwörterbuch der Kriminologie*, hrsg. von R. Sieverts, Band I, Berlin: De Gruyter, 205-231.
- Undeutsch, U. 1967. Beurteilung der Glaubhaftigkeit von Zeugenaussagen. *Handbuch der Psychologie*, Band 11, Göttingen: Hogrefe, 26-181.
- Vernon, Ph. E. 1957. Personality tests and assessment. London: Methuen.

- Wagner, R. 1949. The employment interview: A critical summary. *Personnel Psychology*, 2, 17-46.
- Walther, E. H. 1941. Die Exploration als Mittel zur Handlungsuntersuchung. *Wehrpsychologische Mitteilungen*, 3/2, 23-28.
- Wellek, A. 1958. Exploration und ganzheitliches Verfahren. *Psychologische Rundschau*, 9, 24-28.
- Welzel, U. 1976. Die Rückfallprognose bei Trunkenheitstätern. Faktor Mensch im Verkehr, Heft 25, Darmstadt: Tetzlaff.
- Welzel, U. 1982. Differentielle Prognosekriterien zur Beurteilung der Rückfallwahrscheinlichkeit in das Delikt „Trunkenheit am Steuer“. In: Winkler, W. (Hrsg.): *Verkehrspsychologische Beiträge I. Faktor Mensch im Verkehr*, Heft 32. Braunschweig: Rot-Gelb-Grün Verlag, im Druck.
- White, W. F. 1955. *Street corner Society*. Rev. Ed., Chicago: University of Chicago Press.
- Wiley, L. & Jenkins, W. S. 1964. Selecting competent raters. *Journal of Applied Psychology*, 48, 215-217.
- Wohlfart, E. 1939. Die Bildpostkartenwahl. *Wehrpsychologische Mitteilungen*, H. 5 und 7.
- Wolfe, L. A. 1970. A comparison of the effectiveness of teaching basic interviewing skills by three group methods: Modeling with role playing, role playing, and discussion. *Dissertation Abstracts*, 31 (3-A), 1029-1030.
- Wright, O. R. 1969. Summary of research on the selection interview since 1964. *Personnel Psychology*, 22, 391-413.

8. Kapitel

Biographische Methode und Einzelfallanalyse

Hans Thomae und Franz Petermann

1. Einführung

In einer Einführung zu einer „Soziologie des Lebenslaufs“ hat Kohli (1978) als Grund der „hohen Erwartungen“, die man in den Sozialwissenschaften mit Begriffen wie „Lebenslauf“ und „Biographie“ verbindet, u.a. die Hoffnung genannt, damit den „wirklichen Lebensverhältnissen“ näher zu kommen „als mit den Abstraktionen und Objektivationen der geläufigen Theorien und Methoden“ (Kohli, 1978. S. 9). Diese Erwartungen wurden schon in den frühesten Erwähnungen dieses Ansatzes zum Ausdruck gebracht. Herder (1778) verwies auf die Bedeutung der Erfassung von Lebensläufen in einer Auseinandersetzung mit der als rationalistisch empfundenen Psychologie der Aufklärung. Die empirische Basis einer Psychologie, welche die Ursache menschlichen Verhaltens eher „unter dem Zwerchfell“ als „im Kopf“ sucht, wird in „Lebensbeschreibungen, Bemerkungen der Ärzte und Freunde, Weissagungen“ gesehen. Mehr als ein Jahrhundert später entwarf der Philosoph Wilhelm Dilthey als Gegenbewegung gegen die von W. Wundt, C. Stumpf und H. Ebbinghaus konzipierte „naturwissenschaftliche“ Psychologie in einer bedeutenden Rede vor der Preussischen Akademie der Wissenschaften eine „Verstehende Psychologie“, die ihre Norm in der „Darstellung des Singularen“ finde, so wie sie in der Biographie gepflegt werde. Das „entfaltete seelische Leben“ in seiner Totalität ist ihm die Einheit, von der Aussagen über die menschliche Natur auszugehen haben, nicht fiktive Elemente wie Lust, Unlust, Empfindung, Gefühl. An dieser gelebten Einheit hebt die beschreibende Kunst des Psychologen bestimmte Seiten hervor, wobei die Betrachtung von Lebensabläufen in ihrer jeweils gewordenen Gestalt die sicherste empirische Grundlage abgibt.

Freilich liegt die eigentliche Bedeutung von Dilthey nicht in der (von ihm nur angeregten) Ausarbeitung einer biographischen Methodik als eines wissenschaftlichen Instruments, sondern in seinem unermüdlichen Hinweis auf die „mächtige inhaltliche Wirklichkeit des Seelenlebens“, welche über die der „konstruktiven“ Psychologie zugänglichen Bereiche hinausgehe (Dilthey,

1924, S. 144). Wichtig wird dabei auch die Berücksichtigung des „vollen“ statt des durch methodische Zurüstungen „reduzierten“ Menschen. Deshalb werden von Dilthey besonders geisteswissenschaftliche Epochen analysiert, welche diese Einheit von theoretischer und praktischer Bewältigung der Wirklichkeit besaßen, wie etwa das Zeitalter der Renaissance, das der „französischen Moralisten“ des 17. und 18. Jahrhunderts. Gemäß der These, daß man den seelischen Strukturzusammenhang nur von seinem „voll entwickelten Zustand aus“ erfassen könne, ist die Orientierung an den differenziertesten Erlebnisstrukturen notwendig, wenn man der Eigenart der menschlichen Natur gerecht werden möchte. Statt einer Eliminierung aller Komplikationen zugunsten der Herstellung gewisser laboratoriumsgebundener Standardsituationen fordert die Erfassung des „vollen Menschen“ eine Erweiterung der Beobachtungsbasis gerade in den Bereich des Genialen, des Voraussetzungsvollen, des Höchststrukturierten hinein. Nur hier werden dieser Ansicht gemäß die Grundlinien menschlicher Existenz in prägnanter Weise erfaßt.

Trotz dieser zahlreichen Hinweise auf die Notwendigkeit einer Einbeziehung möglichst umfassender und höchst strukturierter biographischer Einheiten in die wissenschaftliche Diskussion ist es in der auf Dilthey zurückgehenden „geisteswissenschaftlichen Psychologie“ nicht zur Ausbildung einer biographischen Methode gekommen. Wohl hat Misch (1907/62) die Hinweise seines Lehrers in seiner großen Geschichte der Auto-Biographie nutzbar zu machen versucht. Spranger (1966) zog in seiner „Psychologie des Jugendalters“ zahlreiche autobiographische Dokumente, Tagebücher und Gedichte von Jugendlichen zur Veranschaulichung des Ausgesagten mit heran. Auch seine Ansätze zu einer Psychologie des Alters nehmen auf autobiographische Äußerungen von Goethe Bezug (Spranger, 1963). Nirgends geschieht die Auswertung solcher Dokumente aber systematisch. Sie ist auch nicht als Beweisführung gedacht.

In gewisser Hinsicht kann man vielleicht sagen, der Einfluß von Dilthey bezüglich der Entwicklung einer biographischen Methode sei mehr indirekter Natur gewesen. So findet sich bei dem Psychiater H. W. Gruhle mancher Ansatz zur Ausbildung einer derartigen Methodik - und es war sicher kein Zufall, daß dieser sein psychologisches Werk als eine „Verstehende Psychologie“ bezeichnete. Ausdrücklich aber knüpft L. Binswanger (1942, S. 661) an den Begriff der „enthusiastischen Vertiefung“ als der Grundlage eines echten Zugangs zum Verständnis menschlicher Existenz an. Die spezifische Form der „daseinsanalytischen Biographik“ (Binswanger, 1928) ist ohne diese „enthusiastische Vertiefung“ nicht denkbar.

Kuiper (1965) sucht die psychoanalytische Biographik auf die Kategorien des Verstehens zurückzuführen, wie sie bei Dilthey entwickelt wurden. Wenn Gesemann (1924) bei seinen Versuchen zu einer psychologischen Biographik auch nicht auf Dilthey Bezug nimmt, so könnte man zum mindesten seinen

Versuch über Gogol wohl als Beitrag zur Erfüllung des von Dilthey skizzierten Programms werten.

In der jüngeren Vergangenheit ist der Stellenwert der Biographik stark reduziert worden und erst die erheblichen Probleme von formalen Modellen, die zur Abbildung hochkomplexer psychologischer Fragestellungen herangezogen wurden, eröffnen dem biographischen Ansatz eine neue Chance. So entwickelte sich innerhalb der Klinischen Psychologie in den letzten 10 Jahren eine alternative Methodik, die stark am Wesen der Einzelperson und ihrer biographischen Bezüge interessiert ist (vgl. Barlow & Hersen, 1977; Dukes, 1977; Hersen & Barlow, 1976; Kazdin, 1978; Kratochwill, 1978; Leitenberg, 1977; Petermann, 1982; Rüppell, 1979 u.v.a.). Die Analyse der biographischen Bezüge bei diesen prospektiv gesammelten Daten erfolgt mit Hilfe statistischer Verfahren (vgl. die sogenannte „Einzelfallanalyse“).

Die angedeutete Neuentwicklung im Rahmen der Analyse von Einzelfällen weicht hinsichtlich ihrer grundlegenden Intentionen von der biographischen Methode ab. Die biographische Methode klassischer Prägung basiert auf der querschnittlichen Gegenüberstellung von Dokumenten von verschiedenen Personen, und alle Erkenntnisse resultieren meistens aus dem Vergleich von Personen untereinander. Die Einzelfallanalyse begreift sich als intensiver Ansatz, der auf die intraindividuelle Gegenüberstellung von Dokumenten bzw. Informationen von einer Person in verschiedenen Lebensphasen, Heilungsphasen, Therapiephasen u.ä. abzielt.

Die nachfolgenden Ausführungen beschäftigen sich zunächst mit der biographischen Methode und ihrer Anwendung in verschiedenen Bereichen der psychologischen Forschung (u.a. Persönlichkeitspsychologie, Entwicklungspsychologie, Sozialisationsforschung, Psychoanalyse und Streßforschung). Daran schließt sich die Diskussion um die Objektivität der biographischen Methode an; und anschließend wird auf statistische Ansätze im Rahmen der Analyse biographischer Dokumente eingegangen. In den letzten drei Abschnitten wird der extensive Ansatz (= biographische Methode) und der intensive (= Einzelfallanalyse) gegenübergestellt, und die Einzelfallanalyse als empirischer Weg einer Analyse von langfristigen Verläufen vorgestellt.

2. Idiographische Persönlichkeitspsychologie und biographische Methode

Für die Psychologie ist der von Herder und Dilthey aufgezeigte Weg durch G. W. Allport aufgegriffen worden. In der Neubearbeitung seines 1937 zum ersten Mal erschienenen Buches „Personality“ von 1962 wird den „persönlichen Dokumenten und Fallstudien“ eine wichtige Funktion im Instrumentarium des Persönlichkeitspsychologen zugeschrieben. Als „persönliches Doku-

ment“ wird dabei „jede frei geschriebene oder gesprochene Information, die absichtlich oder unabsichtlich Aufschluß über das Leben des Urhebers gibt“ angesehen. Einzelne Formen solcher „persönlicher Dokumente“ sind Autobiographien, Tagebücher, Briefe, Antworten auf offene (nicht standardisierte) Fragen, wörtliche Aufzeichnungen einschließlich Interviews und „gewisse literarische Kompositionen“ (Allport, 1970, S. 393). Er nennt Motive, welche den Inhalt solcher Lebensberichte verfälschen können. Auch die Fallgeschichten von Psychiatern, klinischen Psychologen, Sozialarbeitern, Personalchefs und Mitgliedern anderer Berufe seien dabei zu nennen. Ein Beispiel für die Verwertung „persönlicher Dokumente“ in Form von Briefen hat Allport (1965) in einer seiner letzten Publikationen geliefert: Er erhielt von einem befreundeten Ehepaar 301 Briefe, die eine als Jenny Jove Masterson bezeichnete Frau von ihrem 58. bis zu ihrem siebzigsten Lebensjahr an dieses Ehepaar schrieb. Dabei wird das Problem der Generalisierung von einer einzigen Fallgeschichte (der Entstehung einer immer stärker isolierten und pessimistischen Lebenseinstellung) aufgeworfen und in der Weise gelöst, daß alle möglichen Interpretationsarten des „Falles“ (existenzpsychologisch, tiefenpsychologisch, eigenschaftszentriert) durchprobiert werden. Gerade weil jeder der theoretischen Ansätze etwas an dem Fall erklärt und doch anderes offen bleibt, präsentiert sich das persönliche Dokument als eine Norm, an der sich jede Psychologie messen muß und an der gesehen, die Möglichkeiten und Grenzen psychologischer Bemühungen einzuschätzen sind.

3. Entwicklungspsychologie und humanistische Psychologie

Unabhängig von den geisteswissenschaftlich-verstehenden „Ursprüngen“ biographischer Methoden ist deren Genese in der Entwicklungspsychologie. Kessen, Haith und Salapatek (1970) sehen die „Babybiographien“, die Eltern über die Entwicklung ihrer Kinder in Form von Tagebüchern verfaßten, als den Ausgangspunkt jeder Kinderpsychologie an und verweisen auf Rousseau als geistigen Urheber des Vorgehens. Sonst wird der deutsche Pädagoge Tiedemann (1787) als erster Autor einer solchen Babybiographie genannt. Die eigentliche Blütezeit dieser Babybiographie begann allerdings erst in der 2. Hälfte des 19. Jahrhunderts mit Beiträgen von Preyer (1882), Taine (1877), Darwin (1877) und Champneys (1881) und setzte sich fort in den Babybiographien von Shinn (1900), Scupin & Scupin (1907), Rasmussen (1931, 1934), von Clara und William Stern (1909) und Charlotte Bühler (1922). Die wissenschaftliche Bedeutung dieser Kinderbiographien wurde von Kessen, Haith und Salapatek (1970) trotz all ihrer Nachteile u.a. mit dem Hinweis auf die Tatsache begründet, daß in ihnen die einzelnen Beschreibungen „durch die kontinuierliche und konsistente Existenz eines anderen Individuums miteinander verbunden seien“ (Kessen et al., S. 299). M. a. W.: In vielen dieser Kinderbiographien werden Veränderungen des Verhaltens eines oder einiger Kinder beschrieben. Sie hat-

ten also unmittelbar zu dem Zugang, was eine Gruppe von Entwicklungspsychologen als Gegenstand ihrer Disziplin definiert: nämlich das Studium des Verhaltens, betrachtet im Kontext des Lebenslaufs einer Person. Während in Kinder- und Jugendpsychologie die „Biographie“ an Bedeutung verlor, griff die beginnende Entwicklungspsychologie des ganzen Lebens die Bemühungen um die Ausbildung einer biographischen Methode wieder auf. Hier ist vor allem Charlotte Bühler zu nennen, deren Arbeit über den „menschlichen Lebenslauf als psychologisches Problem“ in sachlicher wie in methodischer Hinsicht einen in seiner Tragweite noch nicht voll ausgeschöpften Ansatz schuf. Sie ließ einerseits „mit der Methode der Anamnese“ (Bühler, 1933) bei einfachen alten Leuten im Wiener Versorgungshaus „50 Lebensgeschichten“ erheben, andererseits werteten in ihrem Arbeitskreis neben Psychologen, „spezialisierte Fachleute, einzelne Historiker, Literatur- und Kunsthistoriker, Soziologen und Mediziner“ (Bühler, 1933, S. 3) literarische Biographien aus, wobei bei der Auswahl vor allem der Gesichtspunkt der Reichhaltigkeit und Zuverlässigkeit der zugänglichen Quellen bestimmend war.

„Interpretationen von Biographien wurden tunlichst nicht verwendet, sondern zur Verwendung gelangten nach Möglichkeit nur die objektiven Daten des Lebenslaufs und dokumentarisch belegte Äußerungen der betreffenden Persönlichkeit selbst über ihr eigenes Leben“ (Bühler, 1933, S. 3).

Das Ziel der Auswertung bezieht sich einmal auf die „allgemeinen Erscheinungen und die reine formale Struktur derselben“ (Bühler, 1933, S. 6).

„Uns interessiert hier nicht, daß Humboldt, weil er ein Romantiker war, weil er aus der und der Familie stammte, weil er mit Goethe und Schiller in Kontakt kam, weil er ökonomisch so gestellt war, daß er sich ein relativ freies Leben erlauben konnte, weil er die und die Neigungen, die und die Gelegenheiten vorfand, nun diesen einen ganz bestimmten Lebenslauf führte, wie er als dieser Charakter, oder er als Vertreter dieses Zeitalters, Kreises usw. es getan hat. Das heißt, uns interessiert an dieser Stelle weder das Individuelle noch das Typologische als solches und auch nicht seine Genese. Sondern uns interessiert Humboldt nur insoweit, als er eine allgemein menschliche Erfahrung in besonders ausgeprägter Weise erlebt und beschrieben hat, so daß wir vermöge seiner die formale Struktur dieses Phänomens ganz besonders prägnant vor uns haben“ (Bühler, 1933, S. 6).

Eine Methode, welche sich an sich zunächst ganz auf Individuen, deren Schicksale, Verhaltens- und Erlebnisweisen konzentriert, wird hier also Instrument generalisierender Aussagen. Die Abstraktion vom Individuellen trotz intensivster Zuwendung zum Individuellen ist hier das kennzeichnende Merkmal.

Diese Abstraktion wird zunächst über den Aufweis von gemeinsamen Erlebnis- und Verhaltensstrukturen in extrem entgegengesetzten Lebensläufen er-

möglichst. Die Biographie eines 74jährigen Rentners und eines 72jährigen Emeritus der Anatomie zeigen trotz des großen sozialen Unterschieds bemerkenswerte Ähnlichkeiten, „sobald man nur die groben Hauptdimensionen und die Verteilung von Zuwachs und Verlust an Dimensionen betrachtet“ (Bühler, 1933, S. 20).

Expansion in der Phase von Jugend- und mittlerem Erwachsenenalter, Restriktion im höheren Alter erscheinen über den Unterschied der Klassen und Epochen hinweg als Grundrichtung menschlichen Daseinsvollzugs, „als objektive Lebenstendenzen“ (Bühler, 1933, S. 83).

Die Ausweitung dieser Studien erfolgte im Arbeitskreis von C. Bühler in Los Angeles (Bühler & Massarik, 1969), wobei die an den Wiener Fällen erarbeitete Grundstruktur auch auf das amerikanische Fallmaterial anwendbar war.

Allerdings ist jetzt die Lebenslaufpsychologie wichtiges Glied einer humanistischen Psychologie, in welcher die Kategorie des „Sinns“ und der Suche nach „Integration“ zu Maßstäben der Beurteilung werden. Die rein deskriptive und abstrahierende Methode wird durch existentielle und therapeutisch orientierte Wertungen ersetzt.

Diese Wertorientierung ist noch entschiedener ausgeprägt bei Abram Maslow (1950), der sein Konzept der Selbstverwirklichung empirisch durch die Auswahl der Lebensgeschichten von 49 bekannten Persönlichkeiten stützte, die das Kriterium der völligen „psychischen Gesundheit“ erfüllen mußten, um in die Liste aufgenommen zu werden. Die ziemlich unsystematisch erhobenen Biographien dienten hier zur Abstraktion der fünfzehn Eigenschaften der ‚Selfactualizer’s‘ und damit zur Definition der menschlichen Norm, an der jeder einzelne zu messen ist.

Die Entwicklung der humanistischen Psychologie zu einer therapeutischen bzw. weltanschaulichen Sekte ist nicht zuletzt durch diese problematische Verwendung der biographischen Methode bedingt gewesen.

Was bei intensiverer Beschäftigung mit den autobiographischen Dokumenten einer Persönlichkeit immerhin zu leisten ist, hat Schmidt (1977) durch seine psychologische Analyse der Biographie von Beethoven gezeigt. Durch Heranziehung der „Skizzen“ zu den Kompositionen und der Tagebücher, Briefe und der künstlerischen Hinterlassenschaft selbst wird das Problem des „Dissens zwischen Biographie und künstlerischem Dasein“ (S. 334) diskutiert. Dabei werden psychopathologische und psychoanalytische Deutungen ebenso als unzureichend erwiesen wie etwa rein geisteswissenschaftliche. „Den Destruktionszwängen auf der biographischen Seite steht auf der künstlerischen eine anscheinend vollkommen intakte Reife und Souveränität entgegen“ (S. 344). Statt irgendetwas „biologischer“ Reduktionen empfiehlt Schmidt (1977)

vielmehr, die zureichende Beschreibung dieser (Beethovens) Konflikte „als biographisch herbeigeführte Interferenzen mit dem sozial determinierten, künstlerisch gestellten Lebensplan, als gewaltige Stauungen und Stimulanzen, zur Bewältigung undurchschaubarer Sozialkonflikte bestimmte Grenzsituationen herzustellen“ (Schmidt, 1977, S. 344). Letztlich aber erscheine die Tragik des biographischen Lebens als „bedingte dialektische Gegenwelt zu ihrer Aufhebung in der Kunst“ (a.a.O. S. 351). Die an der dialektischen Psychologie orientierte Interpretation von Schmidt bleibt den vorliegenden biographischen und künstlerischen Dokumenten sehr nahe und kann in mancher Hinsicht sogar als Beispiel jener „verstehenden“ Methode angesehen werden, die Dilthey (1894) an sich vorschwebte.

Ein etwas festerer Grund für den Psychologen ist bei der Analyse literarhistorischer Quellen etwa mit Hilfe psychologisch bewährter Auswertungstechniken gegeben. So hat Bellak (1964) eine „Thematische Analyse“ im Sinne von H. A. Murray von 10 Kurzgeschichten von Somerset Maugham vorgenommen und sie als Projektion einer sehr selektiven Weltsicht interpretiert, die Ergebnis der Ausbildung bestimmter Abwehrmechanismen gewesen seien. Sears (1974) analysierte sieben Romane von Mark Twain, die dieser zwischen 1868 und 1908 (d.h. im Lebensalter von 33 bis 73 Jahren) schrieb. Da die Teile einiger dieser Romane in mehrjährigen Abständen geschrieben wurden, ergaben sich insgesamt 15 zeitlich voneinander getrennte Episoden des Phantasieausdrucks des Autors. Mit Hilfe der TAT-Auswertungstechnik ordneten zwei unabhängige Beurteiler (mit 76-94%iger Übereinstimmung) die Analyseneinheiten (Episoden) bestimmten Bedürfnisindikatoren zu. Es wurde eine sehr charakteristische Motiventwicklung erschlossen, die bestimmten Ereignissen im Leben von Mark Twain (wie Heirat, Geburt der Kinder usw.) zugeordnet werden konnten.

Ebenfalls an H. A. Murray orientiert ist eine Analyse des dichterischen Werks von Albert Camus durch R. N. Wilson (1964). Gemäß der Grundposition des Autors stellt Psychologie „das Studium von Lebensläufen“ dar und der Dichter, insbesondere der Romancier, sei einer der wichtigsten Lieferanten von Rohmaterial für eine derartige Psychologie. Gerade dieser Versuch von Wilson (1964) dürfte aber in mancher Hinsicht auf die Kritik des literaturhistorischen Fachmannes stoßen, der im übrigen auch bei den anderen hier erwähnten Versuchen die Frage aufwerfen wird, inwieweit das künstlerische Werk unmittelbar als Ausdruck der Persönlichkeit und ihrer Entwicklung verstanden werden kann oder nicht durch literarische Gestaltungsmerkmale bestimmt ist.

4. Probleme psychoanalytischer Biographie

Die Psychoanalyse von Freud wird von manchen als der eigentliche Beginn biographisch orientierter Methodik in der Psychologie angesehen. Man darf in

der Entwicklung dieses Arbeitsansatzes vielleicht drei Ausgangspunkte nennen: den von Breuer behandelten Fall von Frl. O., die Entdeckungen der infantilen Sexualität und die Phase der „Selbstanalyse“ in der Entwicklung von Freud.

Der Erfolg der Psychokatharsis, wie er bei der Behandlung von Frl. O. durch Breuer vorübergehend erzielt worden war, verwies auf den Zusammenhang zwischen einer momentanen Störung und einem zeitlich zurückliegenden unverarbeiteten Erlebnis. Die Repräsentation des Erlebnisses im gegenwärtigen Bewußtsein schafft die Möglichkeit der „Reinigung“. Auf weiter zurückliegende Epochen des individuellen Lebens wird die Aufmerksamkeit des Therapeuten nach der „Entdeckung“ der infantilen Sexualität gelenkt. Das traumatisierende Ereignis muß danach sehr häufig, wenn nicht immer in einer der Phasen der infantilen psychosexuellen Entwicklung gesucht werden.

Als entscheidender Abschnitt in der Ausbildung der psychoanalytischen biographischen Technik aber sieht Jones die im Jahre 1897 von Freud an sich vorgenommene „Selbstanalyse“ an. Aus dem von Jones ausgewerteten Briefwechsel von Freud mit Fließ geht hervor, Freud habe damals „bei sich selbst die Leidenschaft für seine Mutter und die Eifersucht auf seinen Vater entdeckt und dabei die Überzeugung gewonnen, daß dies ein allgemein menschlicher Zug sei, der einem das Verständnis für die gewaltige Wirkung des Ödipusmythos erschließe“ (Jones, 1960, Bd. 1, S. 380f.).

Die spezifische Form psychoanalytischer Biographik erklärt sich aus dieser Ausgangslage. Die Analyse des Lebenslaufs geschieht (fast) ausschließlich in einer vielschichtig - mit Hilfe von „freiem Einfall“, Traumanalyse und provozierter Erinnerung - vorgenommenen Anamnese. Diese hat das Ziel, Verbindungen zwischen der momentanen, durch eine „Störung“ definierten Situation und einem traumatisierenden Ereignis aufzudecken. Dabei ist davon auszugehen, daß diese Traumatisierung in erster Linie in der frühen Kindheit des Individuums erfolgt. Die Jugendzeit, das jüngere Erwachsenenalter und die sich daran anschließenden Phasen erscheinen wenig oder kaum relevant.

Nicht zuletzt aber hat die biographische Analyse in der Art von Freud die Aufgabe und das Ziel, bei dem jeweils vorzufindenden Fall einen „Modellfall“ von biographischem Verlauf vorzufinden. Es geht nicht allein und nicht so sehr darum, noch völlig unbekannte schädliche Ereignisse in der frühen Kindheit mit der späteren Störung in Verbindung zu bringen. Es kommt vielmehr nur darauf an, in dem jeweils zu studierenden Fall die erneute Bestätigung der Lehre über solche Traumatisierungs-Störungs-Sequenzen wieder aufzufinden und die jeweils gegebene konkrete Variation solcher Ereignisse zu demonstrieren (vgl. die nunmehr über 70 Jahre anhaltende Diskussion zum Fall Schreber: Kitay (1963), Niederland (1963)). Handelt es sich um einen eigenen Fall des psychoanalytischen Autors, so wird dieser schon deswegen das generelle ätio-

logische Konzept befolgen, weil der Fall qua „Fall“ ja in jener Interaktion entstand, in welcher der Analytiker dem Analysierten seine an jenem Konzept orientierten Deutungen anbietet und so lange arbeitet, bis sie beiden Seiten akzeptabel erscheinen (vgl. dazu Schraml, 1965).

Von hier aus gesehen haben die zahllosen orthodoxen Fallanalysen, die bis heute vorliegen, und die auch weiterhin veröffentlicht werden, den Sinn der Bestätigung eines generellen biographischen „Modells“ und des Nachweises seiner relativ geringfügigen Modifikation durch die jeweils gegebenen Umstände des späteren Lebens (vgl. etwa De Boor, 1966).

Deshalb bemüht man sich auch, innerhalb der Freud-Orthodoxie die gegebene Interpretation durch Verweise auf Freud's eigene Arbeiten zu stützen. Dieser orthodoxe Kanon der Interpretation aber wurde durch die Ansätze von Adler, Jung, Anna Freud, Heinz Hartmann, Ernst Kris und viele andere in verschiedenster Weise modifiziert. Teilweise legte man andere „Muster“ des überall zu erwartenden Verlaufs zugrunde, teilweise ließ man dem Analytiker einfach mehr Freiheit in der Auswahl des von ihm vorzufindenden Musters (Nachweise u.a. bei Lampl De Groot (1963); Loch (1966); Loewenstein (1960)). In neueren Darstellungen der psychoanalytischen Methoden wird außerdem immer stärker die Durchmusterung des ganzen Lebens gefordert (dazu Schraml, 1965; Kuiper, 1966). Ein Ausdruck dieser Entwicklung ist insbesondere der Versuch von Erikson (1950), die typische Modellsituation von Konflikten, wie sie innerhalb der früheren Psychoanalyse für die Zeit der Kindheit umschrieben worden war, auch für das mittlere Erwachsenenalter zu definieren. Das soziale und politische Engagement von Erikson (1968) war wohl auch der Grund für eine stark analytische Orientierung der politischen Psychobiographie (Wolfenstein, 1967; Mazlish, 1972; Glad, 1973). Die Zielsetzung ist dabei oft sehr weitreichend. So fordert Edinger (1964, S. 668) von der psychologischen Biographie, daß sie eine ganzheitliche Annäherung an die Dynamik der Persönlichkeitsentwicklung darstelle und dabei früher wie gegenwärtig wirksame Bedingungen vollständig erfasse. Naturgemäß ist die Erfüllung einer solchen Forderung bei der Biographie politischer Persönlichkeiten eher möglich als beim „Durchschnittsmenschen“, da die Mitwelt einfach mehr an Informationen erhält und registriert. Aber die Erfassung der formenden Erfahrungen in der frühen Kindheit, der Persönlichkeitsentwicklung während der Adoleszenz und des Erwachsenenalters stellt doch Anforderungen an die Quellen, die oft schwer erfüllbar sein werden. Psychobiographien in diesem Sinne werden als wertvoll für die Formulierung von Hypothesen über die Interaktionen zwischen der sozialen Struktur und dem „psychischen Mechanismus“ angesehen (Glad, 1973, S. 308). Diese Hypothesen könnten dann u.U. durch systematische Formen der Überprüfung ergänzt werden. Genau so sei aber die Generalisierung von dem einzelnen Fall auf umfassendere politische Phänomene möglich.

5. *Biographische Methode als Instrument der Sozialisationsforschung*

Als ein „Markstein“ für die Entwicklung der biographischen Methode wird von Allport (1942) die Arbeit von Thomas und Znaniecki (1918-20) über die Anpassung der polnischen Emigranten in Westeuropa und Amerika angesehen. Hier wurden (mündlich übermittelte) Autobiographien, Briefe, Zeitungsausschnitte, Gerichtsakten und Akten der Wohlfahrtsbehörden als Basis für die Analyse der Akkulturation bzw. Sozialisation herangezogen. Die Autoren gingen dabei von der Überzeugung aus, daß „persönliche Lebensberichte, die so vollständig wie nur möglich sein sollten, den vollendeten Typ des soziologischen Materials darstellen und daß, wenn die Sozialwissenschaften andere Materialien überhaupt anwenden, dies nur geschieht wegen der praktischen Schwierigkeit, im Moment eine genügende Anzahl von solchen Berichten behandeln zu können, zweitens aber wegen des ungeheuren Arbeitsaufwandes, der notwendig ist für eine adäquate Analyse des persönlichen Materials, das notwendig ist, um das Leben einer soziologischen Gruppe zu charakterisieren“ (a.a.O. S. 1832f). Wenn die Arbeit von Thomas und Znaniecki (1918-20) kaum Nachahmer fand, dann nicht nur wegen des großen Arbeitsaufwandes und der gegen die Repräsentativität des Materials erhobenen Einwände, sondern weil andere Methoden in der Sozialforschung in den Vordergrund traten. Für die Erstellung von Biographien trat mehr und mehr das Interview als Erhebungsquelle in den Vordergrund, wobei neben den auf die Verhaltensweisen des Individuums bezogenen Aufgaben mehr und mehr auch der soziale Kontext und sein Wandel Berücksichtigung fanden. Dies trifft insbesondere für die von Davis & Dollard (1940) erhobenen Biographien von jugendlichen Farbigen aus „dem tiefen Süden“ der Vereinigten Staaten zu, in denen versucht wurde, die „Sozialisation“ des Individuums durch die Gruppe, soziale Schicht und die Kultur, in die es hineingeboren wurde, zu demonstrieren. Den Interpretationsvolumen lieferte eine Legierung von S. Freud und C. L. Hall.

Kardiner (1945) und Lewis (1961) sahen in Autobiographien Instrumente des Kulturvergleichs. Nach Lewis (1961) könne man mit Hilfe der Autobiographie vor allem die Gefahr vermeiden, fremde Kulturen mit einer z.B. westlichen oder US-amerikanischen Brille zu sehen.

Lewis führte diese Überzeugung zu der Entwicklung der Methode der „multiplen Autobiographie“, die er zuletzt in einer Studie über Jesus Sandez und seine vier Kinder zu verwirklichen suchte. Die wortgetreue Wiedergabe der Tonbänder, die er von diesen Explorationen aufnahm und die einen Band von fast 500 Seiten Umfang füllen, scheint ihm zunächst durch ein sozialetisches Motiv gerechtfertigt zu sein: Die Klasse der „Armen“, denen diese Familie zugehört, werde weder in der wissenschaftlichen noch in der künstlerischen

Literatur ausreichend berücksichtigt. Dennoch seien ihre Biographien von einem Reichtum, den die offizielle Literatur oft vernachlässige (Lewis, 1961; vgl. auch Paul, 1979). Neue Aktualität gewann die biographische Methode vor allem auch in der Bildungsforschung (Bartenwerfer & Giesen, 1973) und der Industriepsychologie und -Soziologie (Obst, 1961; Bahrdt, 1975; Osterland, 1973; 1978; Garrison & Muchinsky, 1977; Kohli, 1977; 1978). Osterland (1978) glaubt seinen weder nach Erhebungsmethode noch nach Stichprobe näher charakterisierten „Biographien“ von Industriearbeitern entnehmen zu können, daß die insgesamt sehr ungünstigen Lebensbilanzen und -perspektiven der älteren Industriearbeiter Resultat einer lebensgeschichtlichen Entwicklung seien, welche die Jüngeren noch einholen werde. Weit differenzierter sind dagegen die Rückschlüsse, die aus den im Arbeitskreis von Lehr und Thomae (1958; 1965) und Lehr (1969; 1978) gesammelten ca. 1.900 Biographien von Männern und Frauen der Geburtsjahrgänge 1885-1930, die vorwiegend aus der unteren Mittelschicht stammen, gezogen wurden. Gegenüber der Zuordnung von typischen Konfliktarten zu bestimmten Lebensphasen wurde im autobiographischen Material der kurz vor und kurz nach dem ersten Weltkrieg Geborenen der Einfluß der Zeitgeschichte, der politischen und sozialen Veränderungen auf die Auslösung und Intensivierung bestimmter psychischer Krisen deutlich (Lehr & Thomae, 1965). Ein Vergleich der Biographien von Frauen und Männern der gleichen Kohorten zeigte hinsichtlich des beruflichen Schicksals den Einfluß bestimmter Normen, welche eine Identifikation auch berufstätiger Frauen mit ihrem Beruf erschweren. Andererseits konnte gerade durch die Analyse der in allen Kohorten nach dem gleichen Muster erhobenen Biographien deutlich gemacht werden, wie soziale und historisch bedingte Veränderungen, wie etwa die Angewiesenheit der Wirtschaft auf die Frauen während der Kriege oder während einer Vollbeschäftigung, die Funktion jener Normen deutlich schwächer werden lassen und insofern ein selbstverständliches Hineinwachsen in die weibliche Berufsrolle bei den nach 1925 geborenen Jahrgängen ermöglichen (Lehr, 1969).

Lehr (1978) hat die Biographien von 741 Frauen und 570 Männern der Geburtsjahrgänge 1895 bis 1939 hinsichtlich der erlebten Kontinuität bzw. Diskontinuität im Lebensablauf analysiert und damit entscheidende Argumente gegen universalistische Phasen- oder Stufenmodelle gewonnen. Eine sozialisationstheoretische Interpretation des Lebenslaufs erscheint danach angemessener als eine an der biologischen Entwicklung orientierte. Vor allem aber wurde aus der Analyse des autobiographischen Materials die Bedeutung des subjektiven Erlebens „ureigenster individueller Erfahrungen und Erlebnisse“ erkennbar, „die - unabhängig vom biologischen, sozialen oder auch kalendarischen Alter - eine aktive Auseinandersetzung mit der jeweiligen Lebenssituation herausfordern“ (Lehr, 1978, S. 333). Durch die systematische Anwendung der biographischen Methode werden somit unkritische, aber journalistisch gut vertretbare Schlußfolgerungen aus Einzelbeobachtungen, wie wir ihnen etwa

in dem Schlagwort der ‚mid-life-crisis‘ begegnen, weitgehend als unzulässig erwiesen.

6. Psychologische Streßforschung und biographische Methode

Als weitgehend unentbehrlich ist die biographische Methode für die Analyse der Beziehungen zwischen psychischer, sozialer und somatischer Belastung und der Auswahl der Reaktionen auf diese Belastung.

Seit mehr als 25 Jahren haben wir versucht, die Reaktionen von Menschen, die Belastungssituationen ausgesetzt waren, zu erfassen und sie zu Auslösebedingungen und Resultaten in bezug auf bessere Anpassung in Beziehung zu setzen.

Hambitzer (1962) hat in den späten Fünfziger Jahren Körperbehinderte (Amputierte, Querschnittgelähmte usw.) intensiv nach dem Verlauf ihrer Auseinandersetzung mit ihrem Schicksal befragt. Dabei trat deutlich ein gewisser Verlaufstypus hervor, bei dem zu Beginn z.T. eher evasive Reaktionstendenzen dominierten, die sehr leicht in aggressive übergehen konnten, z.B. wenn man versuchte, sie mehr unter Menschen zu bringen oder wenn berufliche Rückschläge zu verarbeiten waren. Später traten dann mehr und mehr verschiedene Formen von Leistung und Anpassung hervor. Doch gab es viele Unterschiede. In fünf Fällen dominierten evasive Techniken bis zu 10 Jahre hindurch, in zwei über 25 Jahre hinweg. Die von Hambitzer Befragten standen im 25. bis 45. Lebensjahr. Es wird somit deutlich, daß Tendenzen zum Ausweichen und Meiden bedrohlicher Situationen keineswegs ein Charakteristikum des Alters sind, sondern eher ein Problem der Höhe des Belastungsgrades.

Das wurde in einer weiteren, an jüngeren Patienten durchgeführten Studie gezeigt. Es handelte sich dabei um Hämophile, die sich einem Heimselbstbehandlungstraining unterzogen hatten, somit also eine aktive Auseinandersetzung mit ihrer Krankheit gewählt hatten. Dennoch fand Kipnowski (1980) bei ihnen viele Hinweise auf evasive Reaktionen und zwar vor allem im krankheitsbezogenen Erlebnisbereich und in ihren Beziehungen zur außerberuflichen Umwelt. Dennoch verweist ein Vergleich von Erwartungs- und Ist-Werten auf eine überzufällige Repräsentanz von Formen der Auseinandersetzung, die auf Anpassung und Behauptung gerichtet sind. Obwohl der Altersbereich hier noch mehr auf Jugend bis zum mittleren Erwachsenenalter bezogen ist, ergibt sich, daß eher die jüngeren zu evasiven Reaktionen, zu Wahrnehmungsabwehr tendieren. Die gelungene Anpassung gerade der stärker Belasteten zeigt sich darin, daß sie eher zu einer positiven Umdeutung der Lage fähig waren.

Die auf den Verlauf der Erkrankung und der Reaktionen auf sie gerichtete Exploration deutete Zusammenhänge zwischen der Aufnahme der Heimselbstbehandlung und der Auswahl der Reaktionsmuster an. Bei diesen Patienten konnte auf eine Abnahme von depressiven Reaktionen, aber auch auf die Tendenz, sich zufällig bietende Chancen aufzugreifen, geschlossen werden, während Tendenzen zum Bagatellisieren von früher sehr beunruhigenden Symptomen, aber auch die psychophysiologische Reaktionsbildung (Stottern, gastrointestinale Beschwerde usw.) sich zurückbildeten.

Andererseits korrelierte die vom behandelnden Arzt eingeschätzte Kooperationsbereitschaft des Patienten („compliance“) signifikant mit einer häufigeren Repräsentanz der Reaktion „Bejahren“ und einer geringeren psychophysiologischen Reaktionsbildung in der Auseinandersetzung mit der Krankheit und der außerberuflichen Umwelt.

Diese Studie konnte nicht die Hintergründe der Auswahl dieser oder jener Reaktionsform aufweisen, wohl aber die Interaktion zwischen Belastungsgrad, Intervention (Beginn der Heimselbstbehandlung), Grad der Motiviertheit des Patienten (Compliance) und des korrespondierenden psychologischen Reaktionsmusters. In einer von U. Lehr betreuten Arbeit von Scharnweber (1980) über die Auseinandersetzung von Dialysepatienten mit ihrer Situation hoffen wir diese Interaktion noch näher klären zu können. Wir glauben, daß Untersuchungen an Patienten, die sich mit einem so hohen gesundheitlichen Belastungsgrad auseinandersetzen müssen, auch für die Interventionsgerontologie von Bedeutung sein können.

Reaktionen auf ‚Life-stress‘ im Alter

In vielen Äußerungen über psychische Reaktionen auf Belastung im Alter wird auf Apathie, Depression, Angst und Hilflosigkeit als dem dominierenden Syndrom verwiesen. Dieses Altersstereotyp wurde auch von dem Psychologen Seligman (1979) übernommen; die Folge von unabänderlichen, der eigenen Kontrolle entzogenen Verluste, die mit dem Alter verbunden sind, führe unweigerlich zu der Ausbildung eines Verhaltenssyndroms, das er als „erlernte Hilflosigkeit“ umschrieb.

Die systematische Erkundung der Reaktionsformen auf gesundheitliche und/oder ökonomische Belastung im höheren Alter mittels ausführlicher Interviews und biographischer Anamnese zeigt demgegenüber, daß Seligman's an Hunden im Pawlow-Geschirr orientierte Theorie der „erlernten Hilflosigkeit“ bestenfalls auf eine sehr kleine Gruppe von institutionalisierten bzw. auf Familienpflege angewiesenen älteren Patienten anwendbar ist. Erkundet man die Formen der Reaktion auf Belastung in den erwähnten Bereichen bei einer einigermaßen repräsentativen Stichprobe älterer Personen, dann wird deutlich,

daß in der Regel sehr aktive und angepaßte Formen der Auseinandersetzung gewählt werden. Unter dem Einfluß von kognitiven Systemen wie jenem der Überzeugung von der Unveränderlichkeit von Belastungssituationen im Alter können allerdings eher depressive und evasive Verhaltensweisen die Oberhand gewinnen, daneben aber auch aktive Reaktionen wie Widerstand gegen die Ausführung ärztlicher Ratschläge in bezug auf Aktivität, Ernährung, Nikotin- und Alkoholgenuß (vgl. Thomae & Kranzhoff, 1979). Die biographische Verankerung dieser kognitiven Systeme wurde durch die systematische Auswertung von Gesprächsaufzeichnungen aus einer zwölfjährigen Beobachtungszeit bei einer kleinen Stichprobe der Bonner Gerontologischen Längsschnittstudie erkennbar, derzufolge zum Teil schon zehn Jahre vor der Messung der Überzeugung der Unveränderlichkeit von Belastung bei Personen mit überdurchschnittlichen Werten in dieser Skala die Thematik „Bestimmtheit von der Endgültigkeit des eigenen Geschicks“ deutlicher ausgeprägt war als bei jenen mit unterdurchschnittlichen Werten.

7. Das Problem der Objektivität der biographischen Methode

Burgess (1945) hat die biographische Methode mit dem Mikroskop des Biologen verglichen. Die Fallstudie erfülle im sozialwissenschaftlichen Bereich die Aufgaben der Vergrößerung und des Durchdringens zu dem, was unter der Oberfläche des äußerlich Beobachtbaren zutage tritt, wie dies das Mikroskop dem Biologen ermögliche.

Es ist allerdings ein offenes Geheimnis, daß der wissenschaftliche Standard der biographischen Technik in den genannten Disziplinen bzw. in den Richtungen der Disziplinen, in denen sie angewandt wird, starke Unterschiede aufweist. Linton (1945) etwa, der aufgrund seiner Zusammenarbeit mit Kardiner kaum als Gegner der Psychoanalyse angesehen werden kann, stellt fest, daß die meisten psychoanalytischen Falldarstellungen aufgrund subjektiver Stellungnahmen gewonnen seien „und nicht jener Art von Beweis unterworfen werden können, wie sie von Mitarbeitern einer exakten Wissenschaft gefordert werden müßte“.

Eine erstmalige Zusammenstellung dessen, was man von einer wissenschaftlich haltbaren Biographie wie Falldarstellung verlangen müsse, wurde von Jaspers einerseits, von Romein für die geisteswissenschaftlich-historische Biographik andererseits gegeben. Dollard (1935) gab vor 45 Jahren „Kriterien“ für die Darstellung biographischen Materials, die Allport und andere einer Revision unterzogen (vgl. u.a. De Waele, 1974; Dailey, 1972; 1975). Faßt man die Thesen dieser Autoren mit eigenen Erfahrungen zusammen, so ergibt sich eine Reihe relativ umschreibbarer Forderungen an den Bearbeiter einer Lebensgeschichte, sofern diese einer Fragestellung in einer der anthropologischen Wis-

senschaften dienen soll. Diese Forderungen sind zum Teil von vornherein als nicht restlos erfüllbar zu bezeichnen. Dennoch müssen sie als Ziel ständig gegenwärtig sein und in möglichst großer Annäherung zu erreichen versucht werden.

1. Die Forderung nach Überschaubarkeit der Bedingungen, unter denen ein berichtetes Phänomen und der Bericht darüber zustande kamen, ist ein Gegenstück zu der Forderung nach Kontrollierbarkeit und Variierbarkeit der Bedingungen eines Versuchsablaufs, die etwa die experimentelle Psychologie stellt.

2. Die Forderung nach Unvoreingenommenheit ist eine wesentliche Vorbedingung der eben erwähnten Vergleichbarkeit von Untersuchungen verschiedener Autoren zum gleichen Thema, es sei denn, man setzt voraus, daß jeder Vergleichende die eigenen Vorannahmen ohne weiteres teile. Dies scheint weitgehend bei allen tiefenpsychologischen und psychoanalytischen Falldarstellungen so zu sein, die dem ohne weiteres stimmig erscheinen, der das Vokabular und die hauptsächlichen Theorien übernahm, die aber dem oft genug absurd erscheinen müssen, der die von dem ursprünglichen Beobachter gestellten „Grundannahmen“ nicht anerkennen kann.

Im übrigen gibt es Behinderungen der Unvoreingenommenheit, welche nicht durch theoretische Einflüsse, sondern durch die persönliche, soziale und berufliche Interessiertheit des Berichtenden an gewissen Endergebnissen bedingt sind. Sie schränken z.B. den Wert fast jeder Autobiographie ein, so unentbehrlich diese für die Bearbeitung soundso vieler Fragestellungen sein mag. Die Bedingungen der Selbstauffassung und Selbsterkenntnis, die jede Autobiographie nur zu leicht zum Mittel der Erhöhung, Bemitleidung, Rechtfertigung, Verteidigung oder Verklärung des eigenen Selbst werden lassen, gestatten eine Verwendung der Autobiographie unter besonderen Kautelen, zu denen etwa die eingehende Beurteilung des Berichters hinsichtlich seiner Fähigkeit gegenüber sich selbst gehört.

3. Die Forderung nach Konkretheit der Aussagen wird nahegelegt angesichts vieler Argumentationen geisteswissenschaftlicher Psychologen, Soziologen oder Philosophen, sofern sie sich zum Beweis für diese oder jene These auf einen „Fall“ oder eine „Lebensgeschichte“ berufen, ohne sich aber um eine nähere Kennzeichnung der Begebenheit - und schon gar nicht um eine solche in soziologischer, historischer oder charakterologischer Hinsicht - zu bemühen. Solche auf keinen Fall zu entbehrenden Hinweise auf die soziologische Einbettung eines Lebensablaufes oder auf seine Färbung durch eine spezifische Temperamentslage und seine Determination durch eine spezifische historische Gegebenheit bedürfen im übrigen durchaus nicht langer Worte, sondern können in wenigen knappen Sätzen gegeben werden.

4. Die Forderung nach Vollständigkeit der darzustellenden Lebensgeschichte

ist wie alle hier erhobenen eine ideale, welche nur in mehr oder minder großer Annäherung erfüllt werden kann. Sie kann sich nur beziehen auf die Verwertung all dessen, was an Daten und Materialien über einen Bios erreichbar ist. „Es gibt keinen Befund, der nicht zur Biographie gehörte und keinen, bei dem nicht sein Ort in der Zeit relevant wäre und sei es sein Charakter der Dauer durch ein Leben.“ (Jaspers, 1962, S. 563). Der Hinweis auf die Wichtigkeit des Details findet sich bei allen historischen Biographien, von Plutarch über Boswell bis zu Romein und Andre Maurois (vgl. Romein).

Diese Forderung nach Vollständigkeit ist freilich in ihrer praktischen Anwendung relativ zu nehmen, d.h. auf die Fülle des gebotenen Materials und den zu behandelnden Stoff bzw. die Art des zu behandelnden Problems zu beziehen. Welche Folgen eine lückenlose Aufzählung aller Verhaltensweisen - wäre sie überhaupt zu erreichen - für den Umfang von „Fallstudien“ hätte, mag der Hinweis auf die sorgfältigen und den Kinderpsychologen noch heute unentbehrlichen Beobachtungen von Jaehner (1930) über zwei Tage aus dem Leben zweier Geschwister zeigen: Die einfache Wiedergabe der beobachteten Situationen und Verhaltensweisen ergab ohne Kommentar 113 z.T. eng bedruckte Seiten (vgl. auch Barker & Wright, 1955). Dort, wo infolge ähnlich reichlich fließender Quellen nicht alles in die Niederschrift aufgenommen werden kann, empfiehlt es sich, vorher einen Verhaltenskatalog anzulegen, der in das Schema der wichtigsten Lebensabschnitte des zu beschreibenden Individuums eingetragen wird. Wesentlich ist dabei, möglichst ein Bild aller Bezüge des Dargestellten zu erhalten und sowohl jene oft schon selbstverständliche Zentrierung der Reaktionen um die Sphäre von Genuß- und Behauptungswerten zu vermeiden, wie sie nicht nur die psychoanalytischen Arbeiten kennzeichnet, als auch jene Lebensbilder aus der Perspektive engherzig gewordener Fürsorgebeamten, auf die sich viele Schilderungen von sozial auffällig gewordenen Persönlichkeiten - auch in der gelehrtesten psychopathologischen Literatur - beschränken.

Biographische Darstellungen wie im übrigen jede charakterologische Schilderung müssen Vollständigkeit auch in dem Sinne anstreben, daß sie nicht nur die von bestimmten sozialen Normen aus festzustellenden Mängel des beschriebenen Menschen registrieren, sondern auch positive Aussagen darüber machen, wie einem Menschen sich das Dasein von innen gesehen möglich macht.

Unsere wiederholten Einwände gegenüber der Psychoanalyse wollen im übrigen nicht darüber täuschen, daß gerade aus ihrem Kreise die eindeutigsten Angaben über Kriterien der Vollständigkeit einer Biographie kamen, und zwar zum einen von Dollard, zum andern von Kardiner und Murray.

Dollard nennt sieben Punkte, die beachtet werden müssen, damit eine „life-history“ als vollständig bezeichnet werden könne. Zu ihnen gehört:

1. Die „Betrachtung des Subjekts als ein Exemplar in einer kulturellen Reihe“,
2. die Anerkennung der sozialen Bedeutung der biologischen Motivation,
3. die Berücksichtigung der Rolle der Familie bei der Übermittlung des zivilisatorischen Standpunktes,
4. der Aufweis der Art der Verarbeitung von biologischen Faktoren im sozialen Verhalten,
5. die Wiedergabe möglichst aller für das Individuum bedeutsamen Eindrücke vom Kindes- bis zum Erwachsenenalter,
6. die sorgfältige und kontinuierliche Spezifizierung der sozialen Situation,
7. die begriffliche Einordnung des Materials der life-history.

Dollard glaubt an Hand einer Gegenüberstellung je einer Falldarstellung von Adler, Taft, Freud, Thomas und Znaniecki, Clifford, R. Shaw und H. G. Wells zu zeigen, daß Freud diesen sieben Punkten und damit den Erfordernissen einer Vollständigkeit der Biographie am meisten gerecht wird. Es steht nun freilich dahin, ob man aus dieser Feststellung - wie Allport dies tut - folgern muß, die Prinzipien von Dollard seien eben nichts als angewandter Freudianismus und entsprächen keinen realen Erfordernissen. Immerhin enthalten die Kriterien von Dollard in (1), (3), (5) und (6) Forderungen, wie sie die gesamte sozialanthropologische Schule von Malinowski bis zu Kimball Young vertritt, nämlich die genaue Spezifizierung der kulturellen, soziologischen und ökonomischen Faktoren, welche die Entwicklung des Individuums beeinflussen konnten. Dieser Forderung ist in den meisten deutschen Falldarstellungen wenig Rechnung getragen worden, da hier konstitutionelle Faktoren und nicht situative im Mittelpunkt der Beachtung stehen. Sucht man die hier bereits erörterten Gesichtspunkte einer Beurteilung der Vollständigkeit von Lebensgeschichten zusammen mit den nicht eigens explizierten von Kardiner und Murray auf einige wenige, dafür aber unbedingt verpflichtende Punkte zu reduzieren, so läßt sich sagen: die Lebensgeschichte (Falldarstellung) muß

- a) den kulturellen, soziologischen und ökonomischen Rahmen skizzieren, in dem sich ein Bios vollzieht,
- b) sie muß jeweils festzustellen suchen, wieviel von diesem Rahmen subjektiv bedeutsam wird und wieviel nicht,
- c) sie muß die konstanten Merkmale einer Persönlichkeit in den verschiedenen Lebensabschnitten (wie etwa die Größe und Richtung des Antriebs, Differenzierungs- und Strukturierungsgrad, Intro- bzw. Extraversion, Art der Grundstimmung, Steuerung unter Angabe des führenden dynamischen Kerngebietes) festhalten,
- d) sie muß die Varianten des Verhaltens in den verschiedenen Lebensepochen möglichst sorgfältig zu erkennen geben, also die meist nur schwer zugänglichen Veränderungen und Wandlungen im Persönlichkeitsgefüge, wie sie etwa in den Begriffen Verfestigung, Erstarrung, Lockerung, Vertiefung, Verflachung, Verinnerlichung, Distanzierung, Vergrößerung, Versandung usw. zutage treten;
- e) sie muß den zu betrachtenden Bios nicht nur von bestimmten sozialen Normen, sondern auch von den für ihn wesentlichen Anliegen aus zu erfassen suchen. Insbesondere muß in einer Biographie ebenso wie in der kleinen Falldarstellung erkennbar

werden, wie ein Mensch sich das Dasein möglich zu machen sucht und nicht nur, was die Sozietät an ihm vermißt bzw. auszusetzen hat.

Bei vielen literaturhistorischen und politisch-historischen Biographien bedurfte es sehr umfangreicher historischer Fachkenntnisse, um die Qualität der Quellen einer Biographie beurteilen zu können. Deshalb wird ein systematischer Gebrauch solcher Biographien erst durch eine interdisziplinäre Zusammenarbeit ermöglicht werden. Diese ist im Augenblick kaum gegeben.

Gegenüber der Intensität der Bemühungen um größtmögliche Objektivität der biographischen Methode, wie sie in den Jahren ihrer Entstehung zu bemerken waren, müssen neuere Versuche (z.B. Friedrichs & Kamp, 1978) als unzureichend angesehen werden. Zur Belebung der aktuellen Diskussion und zur Abklärung der Position der biographischen Methode muß diese mit dem experimentellen Vorgehen kritisch verglichen werden. Es wäre auch denkbar, daß auf diesem Hintergrund neue Kriterien an die biographische Methode gestellt werden müssen, die die hier aufgeführten ergänzen.

Im Rahmen der Gegenüberstellung von biographischer Methode und dem experimentellen Vorgehen sollen die aufgeführten Forderungen und deren Erfüllbarkeit mit den Ansprüchen des psychologischen Experiments verglichen werden. Zur Strukturierung dieser Gegenüberstellung werden in Tabelle 1 die entsprechenden Kriterien aufgelistet.

Tabelle 1: Gegenüberstellung von Kriterien, die an die Biographik und das experimentelle Vorgehen angelegt werden müssen.

Kriterien an die Biographik	Kriterien an das experimentelle Vorgehen
Überschaubarkeit der Bedingungen	Kontrollierbarkeit und Variierbarkeit der Bedingungen
Unvoreingenommenheit des Beobachters	Vergleichbarkeit von Untersuchungsergebnissen aus verschiedenen Experimenten
Konkretheit der Aussagen	Präzision der realisierten experimentellen Bedingungen, die die Objektivität gewährleisten
Vollständigkeit der darzustellenden Lebensgeschichte	Repräsentativität der experimentellen Bedingungen für die Abbildung psychischer Realität

In den vorangegangenen Ausführungen wurde schon die Forderung 1 mit der nach der Kontrollierbarkeit und Variierbarkeit der Bedingungen des Experi-

ments in Beziehung gesetzt; inwieweit diese Forderungen bzw. Kriterien an experimentelles Arbeiten dann sinnvoll angelegt werden können, wenn dieses psychologische Realität abbilden soll, muß kritisch beurteilt werden (vgl. McGuire, 1973). Die Forderung 2 (Unvoreingenommenheit des Beobachters) ist sicherlich genauso schwer zu erfüllen, wie Vergleichbarkeit von Untersuchungsergebnissen aus verschiedenen Experimenten. Ein Vergleich der Forderung 3 (Konkretheit der Aussagen) und 4 (Vollständigkeit der darzustellenden Lebensgeschichte) mit Kriterien des experimentellen Vorgehens zeigt ähnliche Ergebnisse. Die Präzision der realisierten experimentellen Bedingungen sinkt immer dann entscheidend, wenn psychologische Realität komplex abgebildet wird oder spezifische Fragestellungen (z.B. aus dem Bereich der Veränderungsmessung oder der Kausalanalyse) zur Lösung anstehen. Es ergeben sich z.B. im Bereich der Veränderungsmessung eine solche Anzahl von Schwierigkeiten (vgl. Petermann, 1978), daß bei experimentellen Ansätzen erhebliche Interpretationsprobleme auftreten. Den Fragen nach der Repräsentativität (vgl. externe Validität bei Campbell & Stanley, 1970) der experimentellen Bedingungen für die Abbildung psychischer Realität wird zwar sehr große Bedeutung eingeräumt, jedoch sind die Bemühungen zu ihrer Erlangung noch sehr rudimentär (vgl. Kirchner et al., 1977).

Die Gegenüberstellung weist keine der Vorgehensweisen als eindeutig überlegen aus, wenn man die Anzahl der Probleme zur Erreichung der Idealform zugrundelegt. Die Präferenzierung bestimmter Erkenntnismethoden scheint daher eher in der Verwurzelung bestimmter Forschungstraditionen als in ihrer unbestreitbaren Vorteilhaftigkeit zu liegen.

8. Vorschläge zur Erhöhung der Objektivität der biographischen Methode

Wie schon angedeutet, besteht eine Möglichkeit der Erhöhung der Objektivität der biographischen Methode darin, die ausgearbeiteten Forderungen weiter zu präzisieren. Eine solche Präzisierung scheint uns zumindest in zweifacher Weise möglich zu sein:

- (1) Man sollte die dargestellten Forderungen durch das Kriterium der Dokumentierbarkeit, d.h. des Festhaltens von biographischen Informationen nach expliziten Kriterien ergänzen, so daß die wissenschaftliche Nachvollziehbarkeit und damit die Glaubwürdigkeit dadurch gegeben ist und Dritte die Dokumentationsregeln erkennen können. Anders formuliert: Das Ausmaß an Objektivität der biographischen Methode wird durch die Erfüllung von apriori festgelegten Dokumentationsregeln realisiert.
- (2) Als weiteres Kriterium könnte man zur Eingrenzung der Forderung nach der „Vollständigkeit der darzustellenden Lebensgeschichte“ die Zentralität

von Ereignissen und Merkmalen in der Lebensgeschichte bezogen auf die jeweilige Fragestellung formulieren. Dieses Kriterium erscheint notwendig auf dem Hintergrund der unvermeidlich vorzunehmenden Reduktion des vorfindbaren Bedingungsgefüges.

Die praktische Umsetzung dieser Kriterien muß Konsequenzen für die Art der Datenerhebung haben. Als wesentliches Problem ergibt sich dabei die schrittweise Erhöhung der Urteilssicherheit bei der Verwendung von biographischen (= retrospektiven) Daten. Eine hohe Urteilssicherheit bei biographischen Daten ist dann erreicht, wenn verschiedene Informationsquellen in einem Gesamturteil verbunden (verschränkt) werden und zudem die verschiedenen Informationsquellen (vgl. Abb. 1) zu denselben Schlüssen führen.

Zur Illustration des Vorgehens der schrittweisen Erhöhung der Objektivität von biographischen Daten soll Abbildung 1 herangezogen und im folgenden beschrieben werden. Die Abbildung geht von vier Schritten der Erhöhung der Objektivität aus, die hinsichtlich ihrer Zielsetzung, ihres Vorgehens und der notwendigen Hilfsmittel untergliedert werden.

Für die Zielvorstellungen ist entscheidend, daß bei jedem Schritt durch eine andersgeartete Urteilsverschränkung die Rekonstruktionsregeln der individuellen Lebensgeschichte validiert werden. Im Detail erscheinen die Aufdeckung von selbstbezogenen Merkmalen (Selbstwahrnehmung, Körpergefühl, Selbstkonzept u.ä.), die Wechselwirkung von selbstbezogenen Merkmalen, die interpersonellen Bezüge und die dokumentierten externen Bezüge bedeutsam. Für die Rekonstruktion individueller Biographien ist diese mehrschichtige Vernetzung zu verschiedenen „Zeiten“ (bezogen auf vorangegangene Lebensphasen) relevant.

Das Vorgehen und die gewählten Hilfsmittel zielen auf die Präzisierung der Problemsituation in der Vergangenheit ab, die sich auf personale und situationale Aspekte beziehen (vgl. ein Verfahrensvorschlag aus der Kinderpsychotherapie von Petermann & Petermann, 1980). Da physikalische und subjektive Zeitrasterung nicht identisch sind, d.h. bestimmte Krisen-/Umbruchphasen werden bedeutsamer und zeitlich sich länger erstreckend wahrgenommen als andere, müssen zur Aufschlüsselung der Erlebnisdichte die Massierung von positiven und negativen Erlebnissen sowie Hoch- und Tiefpunkte exploriert werden (Abschnitt 10).

Vielfach wird es für die Versuchsperson eine Erleichterung bedeuten, ihr Auf- und-Ab in ihrer Biographie in einem Verlaufsdiagramm abzutragen und die Massierung im Erlebnissbereich zu kennzeichnen (vgl. dazu Stimmungsdiagramme der Berkely-Growth-Study; s. a. Bühler & Massarik, 1969). Auf diese Weise lassen sich die Hoch-/Tiefpunkte in verschiedenen „Zeiten“ vergleichen; verschiedene Zeiten (Phasen) lassen sich durch Rollenwechsel, Veränderungen

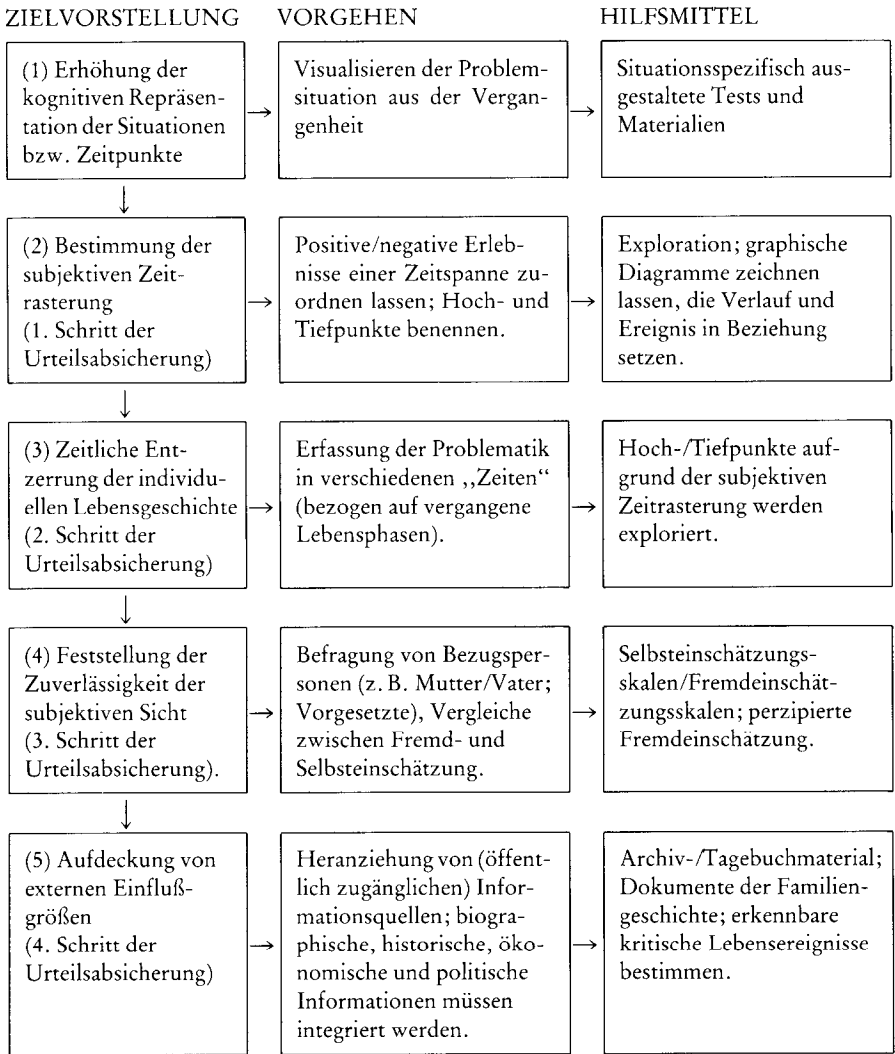


Abb. 1: Schritte zur Erhöhung der Objektivität von biographischen Daten (in Anlehnung an Petermann, 1982, S. 48).

im Erleben von Abhängigkeit und Macht u.ä. definieren. Die Selbsteinschätzungen lassen sich durch die Befragung von Bezugspersonen (3. Schritt in Abbildung 1) und die Heranziehung von Informationsquellen (historische, ökonomische und politische Dokumente; vgl. 4. Schritt in Abbildung 1) objektivieren.

9. Das halbstrukturierte Interview und das Problem der Kontrolle der Datengewinnung

Auch bei den meisten „persönlichen Dokumenten“ wie Tagebüchern, Briefen, Autobiographien ist die Möglichkeit eingeschränkt, die Validität der Aussagen abzuschätzen, da man die Motivationen und darstellerischen Zielsetzungen der Autoren nicht erfassen kann.

So bleibt für die Gewinnung von Informationen über die psychologische Biographie eines Individuums das unter standardisierten Bedingungen durchgeführte halbstrukturierte Interview oder die Exploration. Wenn man „Interview“ und den aus der Psychiatrie stammenden Begriff „Exploration“ voneinander abhebt, so hat dies keineswegs nur status- oder einstellungsbezogene Gründe. Bildet für das Interview die sozialwissenschaftliche Enquete das Paradigma schlechthin, so ist für die (psychiatrische) Exploration die Herstellung einer spezifischen sozialen Beziehung Voraussetzung. Je mehr an Äußerungen von dem Befragten erwartet werden und je zentraler diese Äußerungen dessen ureigenste Lebenssphäre berühren, desto mehr muß sich der Befragte wirklich als „Mensch“ angesprochen fühlen. Dies aber ist offensichtlich nur zu erreichen, wenn sich der Untersucher selbst in ausreichendem Maße engagiert. Dies setzt wiederum die Sicherung einer sozialen Situation während des ganzen Gesprächs voraus, bei der die Rolle des „Gebenden“ und des „Nehmenden“ in einer Weise wechseln, wie dies dem „natürlichen“ Gesprächsverlauf entspricht. Dies darf natürlich nicht dazu führen, daß die Richtung entscheidender Antworten beeinflusst wird oder daß der Befragte daran gehindert wird, einen spontanen Bericht über den „Grundriß“ seines Lebens, über ihn besonders interessierende Episoden aus diesem Leben oder über besonders relevante Einheiten - z.B. den Tageslauf - möglichst zusammenhängend zu geben.

Ein wichtiges Prinzip für diese Art der Informationsgewinnung aber beruht in folgendem: Die Exploration soll dem sogenannten „Durchschnittsmenschen“ eine Chance geben, in der Wissenschaft vom menschlichen Verhalten und seiner inneren Begründung zu Wort zu kommen. Die meisten unserer diagnostischen Verfahren engen seine Antwortmöglichkeiten bereits auf ein Konzept ein, das den Erwartungen einer bestimmten Theorie oder den Erfordernissen einer bestimmten Methodologie entspricht. Dadurch verschließt sich die Wissenschaft den Zugang zur vollen Breite menschlichen Verhaltens. Da eine Fremdbeobachtung dieses Verhaltens aus äußeren Gründen meist nicht möglich ist, stellt die Exploration einen der wenigen Zugänge zu einer durch den methodologischen Zugriff noch nicht veränderten seelischen Wirklichkeit dar. Insofern möchten wir die Exploration als jene Form einer Interaktion zwischen „Untersucher“ und „Befragten“ charakterisieren, bei der der Untersucher ganz auf den Status des „überlegenen“, des „Experten“, dessen, der

Dinge durchschaut, die der andere nicht sieht, verzichtet und den Partner wirklich als Partner anerkennt (vgl. dazu auch Mierke, 1944; Kinsey et al., 1948).

Als Vorbedingungen der Interaktion, die der Gewinnung einer Autobiographie dienen soll, ist somit eine emotional und affektiv weitgehend „entlastete“, dabei aber doch zum „Engagement“ auffordernde Situation anzusehen. Insofern bedarf es jeweils einer „Motivierung“ des Gesprächs, welche das Interesse des Partners zu wecken und zu erhalten vermag. Im Gegensatz zu vielen Themen der Sozialforschung, aber auch der „Motivforschung“ ist das Wecken dieses Interesses für persönlichkeitspsychologisch relevante Fragen oft ohne größere Vorbereitungen zu erzielen (Whyte, 1955; Bain, 1960). Man kann Jugendliche durchaus dazu gewinnen, über die „wahren“ Hintergründe ihrer Konflikte mit der älteren Generation zu sprechen und von hieraus zu einer Darstellung des bisherigen Lebensablaufs zu bewegen. Ebenso sind ältere Menschen durchaus dafür zu interessieren, zur Erkenntnis der „Schwierigkeiten“ ihrer Lebens Epoche beizutragen. Ebenso ist es möglich, den Hochofenarbeiter zum „Reden“ zu bringen, wenn er das Gefühl erhält, „denen“ zeigen zu können, wie es wirklich mit dem Arbeiter bestellt ist. In all diesen Fällen schwingt als Motiv die Genugtuung darüber mit, daß andere sich für das eigene Schicksal interessieren, nicht selten aber auch die Bereitschaft, anderen Menschen durch den Bericht über die eigene Lebenssituation behilflich zu sein. Richardson, Dohrenwend und Klein (1965) nennen „Altruismus“, „Befriedigung über die Möglichkeit, sich auszusprechen“ und „intellektuelle Genugtuung“ als wichtigste Ansatzpunkte für die Vorbereitung einer psychologischen Exploration, die zu Forschungszwecken durchgeführt wird. Auch Kinsey und seine Mitarbeiter (1948) verließen sich in hohem Maße auf diese Motivationskomplexe.

10. Die Frage der „Einheiten“ der Biographie

Die Herausarbeitung der „Einheiten“ eines Lebenslaufs ist zwar letztlich Aufgabe der Auswertung und der systematischen Analyse. Für die Planung der Explorationsmethode ist jedoch die Berücksichtigung des Problems dieser „Einheiten“ so wesentlich, daß schon jetzt darauf einzugehen ist.

Der Untersuchungspartner, der zur Darstellung seiner Biographie aufgefordert wird, wird naturgemäß diese äußerst globalen Einheiten des Lebenslaufes in kleinere Einheiten aufgliedern. Die Besinnung auf die frühe Kindheit wird ohnehin vom Untersucher durch entsprechende Fragen nach „Früherinnerungen“ gefördert werden. Hier wird dann eine Zeit von Jahren durch mehrere Episoden, im übrigen aber durch äußere Daten und vor allem durch das „Bild der Eltern“ beleuchtet. Soweit psychoanalytische Orientierung besteht, wird

man jene „Episoden“ dann durch Deutung von Träumen und „freien Einfällen“ zu erweitern suchen.

Eine deutlich abgrenzbare Einheit stellt dann für viele die Schulzeit und die Zeit der Berufswahl dar, wobei auch hier die Gesamtheit durch kleinste Einheiten (Episoden) beleuchtet wird. Bei den älteren Jahrgängen, insbesondere den Männern, werden sehr oft Kriegszeit und Nachkriegszeit zu bestimmten Gesprächsthemen, die unter Umständen den meisten Raum der Autobiographie ausfüllen können. Dies ist besonders dann der Fall, wenn ausführliche Darstellungen von Erlebnissen im „Einsatz“ oder in der Gefangenschaft sich in den Vordergrund schieben. Bei weiblichen Probanden nimmt häufig die Schilderung der Familiengründung, der Geburten, der Flüchtlings- oder Evakuierungszeit einen entsprechend breiten Raum ein.

Von dieser Eigenheit des spontanen Berichtes über die Autobiographie aus empfiehlt sich eine Anwendung der Serie der standardisierten (aber offenen) Fragen, welche den Blick wechselweise auf die „globale“ Einheit und auf kleinere Einheiten lenkt. So kann die Frage nach dem „hinderlichsten“ Ereignis als Eröffnungsfrage für eine erneute globale Darstellung des Lebenslaufs dienen. Es kann aber auch - je nach der gegebenen Fragestellung - die frühe Kindheit, die Schulzeit, die Zeit der Berufsfindung oder der Berufswahl, die Zeit insbesondere starker ökonomischer und allgemein historischer Belastungen (politische Verfolgungen, Krieg usw.) zum Ausgang einer gezielten Serie von Fragen genommen werden.

Wichtig ist dabei auch, daß stets die Möglichkeit eines Vergleichs besteht zwischen kleinsten biographischen Einheiten, wie sie in Handlungen gegeben sind, mittleren biographischen Einheiten, wie sie in bestimmten „Episoden“, d.h. durch äußere oder innere Veränderungen herbeigeführten Perioden der Umstellung vorliegen und längeren Einheiten von relativ konstantem Verlauf. Nur dann ist eine Chance dafür vorhanden, die später folgende Analyse einerseits genügend konkret, andererseits genügend umfassend vorzunehmen.

Bei der Erkundung der einzelnen Einheiten ist es vor allem wesentlich, Informationen über Verhaltens- und Erlebnisweisen des Berichtenden selbst und nicht nur globale, in Attributen oder Adjektiven zusammengefaßte Charakterisierungen zu erhalten. Denn da eine psychologische Analyse solcher Daten sich auf Verhaltens- und Erlebnisweisen beziehen soll, helfen globale Kennzeichnungen wie „damals war eine glückliche Zeit“ nicht viel.

Von psychiatrischen (Pauleikhoff), soziologischen und psychologischen Erfahrungen aus wird mehrfach der Vorschlag gemacht, zur Ergänzung bzw. als Verbindungsglied der übrigen biographischen Einheiten den selbstberichteten Tageslauf zu wählen. Ein derartiger Bericht kann sich zwar stets nur auf die Gegenwart beziehen, vermag aber hier Informationen über formale Ver-

haltens- und Erlebnisweisen zu liefern, die in engem Zusammenhang mit Verhaltens- und Erlebnisstrukturen stehen, wie sie aus früheren Epochen vorliegen (Olbrich).

Außerdem kann die Erhebung des Tageslaufes in Längsschnittuntersuchungen wenigstens in Abschnitten über 2 Jahre hinweg vorgenommen werden. Die Erkundung des Tageslaufes hat vor allem den Vorteil, daß sie eine Möglichkeit der Erfassung des „Routine“-Verhaltens in seiner ganzen Breite bietet.

Als eine weitere „mittlere“ Einheit der Erfassung sollte man auch mehr und mehr die Analyse des typischen „Jahreslaufes“ ausbauen, da sie ja das Wiederkehrende und das Wechselnde des Verhaltens in noch deutlicherem Licht zeigen kann als die Erfassung der übrigen Einheiten.

Wesentlich bleibt nur als Prinzip der anamnestisch orientierten Exploration das ständige Bemühen um die Verbindung von „Konkretheit“ und „Globalität“ der Aussage. Um dies zu erreichen, ist das Aufsuchen der hier genannten Einheiten notwendig. Es kann im übrigen in beliebigem Maße systematisiert werden.

11. Eine Möglichkeit der statistischen Auswertung von Biographien

Ausgehend von der Krisendiskussion des experimentellen Arbeitens in der Sozialpsychologie und der Forderung von McGuire (1973) verstärkt Archivdaten, persönliche Dokumente, historische Quellen zur Aufdeckung von sozialen Bezügen und persönlichen Eigenschaften zu nutzen, unternahm Simonton seit 1975 den Versuch, mit Hilfe von Klein-N-Studien (Studien mit Stichproben um ca. 20 Versuchspersonen) biographische Daten zeitreihenanalytisch auszuwerten. So betrachtet Simonton (1975) Kreativität in Abhängigkeit vom Entwicklungsverlauf von historischen Persönlichkeiten der europäischen Geschichte und zeigt, daß Kreativität von soziokulturellen Bedingungen, wie der politischen Stabilität, beeinflusst wird.

Eine weitere, sehr umfassende Analyse der Biographien von 301 Genies - basierend auf biographischen Unterlagen von Cox - führte Simonton (1976) durch. Simonton versuchte, den Einfluß des Status des Vaters, der Intelligenz und der Erziehung näher zu bestimmen, wobei er zwischen Führerpersönlichkeiten und kreativen Persönlichkeiten Unterschiede genauer analysierte.

In einer weiteren Studie beschäftigte sich Simonton (1977b) mit der Produktivität von 10 bekannten europäischen Komponisten (Bach, Beethoven, Mozart, Haydn, Brahms, Händel, Debussy, Schubert, Wagner und Chopin); er diskutierte, ob die Produktivität - bezogen auf verschiedene Lebensphasen - durch äußere Faktoren, wie Krankheit, Streß u.ä., zu beeinflussen ist. Unter

Zugrundelegung der verfügbaren Zeitdokumente unterteilte er dabei den Lebenslauf der Komponisten in 5-Jahresperioden, die er hinsichtlich bestimmter Produktivitätsmerkmale (Werke und Themen) untersuchte.

Unter methodischer Sicht sind die Bemühungen Simontons besonders interessant, vor allem, wenn man an eine statistische Analyse von biographischen Daten denkt. Simonton (1977a) schlägt zur statistischen Betrachtung das relativ einfache Modell der Regressionsanalyse (hier in der Form der Regressionsanalyse über zeitliche Verläufe = Zeitreihenanalyse) vor. Obwohl dieses einfache Modell in verschiedenen Ausgaben des *Psychological Bulletin* (1978 bis 1979) kritisch diskutiert, und von Swaminathan & Algina (1977) eine mathematisch umfassendere und präzisere Betrachtung angestellt wurde, soll kurz auf die Logik des Verfahrens eingegangen werden (vgl. Abschnitt 12).

Das Vorgehen von Simonton vergleicht Dokumente in verschiedenen Lebensphasen miteinander (z.B. die Produktivität von Komponisten in 5-Jahresabständen bzw. bestimmten Lebensphasen). Zentral sind dabei Überlegungen hinsichtlich der genaueren Unterteilung der Lebensphasen. Für die Anwendung des statistischen Verfahrens sind zumindest vier Lebensphasen notwendig (vgl. Petermann, 1980). Hinsichtlich jeder Lebensphase und über die Lebensspanne lassen sich für alle Biographien oder eine bestimmte Auswahl (Führerpersönlichkeiten vs. Kreative; Simonton, 1976) bestimmte auf den Entwicklungsverlauf bezogene Hypothesen formulieren und testen. Die zeitreihenanalytische Dokumentenanalyse nach Simonton macht es möglich, angenommene Verläufe bei den zugrundegelegten Biographien und ihre Wechselwirkung zu bestimmten abhängigen Variablen zu untersuchen.

12. Biographik und Einzelfallanalyse

Die Überlegungen zur schrittweisen Erhöhung der Objektivität von biographischen Studien, die sich im wesentlichen durch eine Erhöhung der Informationen über den Einzelfall erreichen läßt, legt eine intensive Beschäftigung mit dem Einzelfall nahe. Die abschließenden Abschnitte möchten daher die intensive, auf den Einzelfall bezogene Analyse von Verläufen als eine mögliche Perspektive des Forschungsgebietes ansprechen. Jaspers (1913) wies schon darauf hin, daß die Erkenntnisbasis der Psychopathologie aus der Analyse des (individuellen) Krankheitsverlaufes sich ergibt, wobei typische Verlaufsreihen hinsichtlich phasenhafter und periodischer Auffälligkeiten (z.B. bei Melancholie oder manisch-depressiven Zuständen) zu untersuchen sind. Für solche Analysen ist die Unterscheidung in abrupte Veränderungen (z.B. bei epileptischen Anfällen) und allmählichen Veränderungen (z.B. beim Wachstum, bei Rückbildung oder Reifung) von großer Bedeutung; ebenso die Analyse des Beginnens und Abklingens von Phasen. Eine von Jaspers geforderte Detailana-

lyse von Verlaufsparametern (abrupte vs. allmähliche Veränderungen; Periodiken u.a.; vgl. auch Petermann, 1982) ist für die Klinische Psychologie von großer Bedeutung, und so ist es auch nicht allzu sehr verwunderlich, daß man in den 70er Jahren in der Klinischen Psychologie damit begann, sich besonders intensiv um die statistische und qualitative Analyse von Einzelfällen zu kümmern. Wie schon einleitend erwähnt, entwickelt sich in diesem Zweig der Psychologie die sogenannte „Einzelfallanalyse“ (Anton, 1978; Baer et al., 1968; Barlow & Hersen 1977; Hersen & Barlow, 1976; Kratochwill, 1978; Leitenberg, 1977; Neufeld, 1977; Petermann, 1977; 1982).

Im Bereich der Klinischen Psychologie ist die Einzelfallanalyse (Einzelfallstudie, N= I-Studie, kontrollierte Fallstudie) weitgehend an das operante Paradigma gekoppelt. In der Verhaltenstherapieforshung wird die Einzelfallanalyse zur Überprüfung von Therapieeffekten eingesetzt, wobei die Zeitverläufe eng eingegrenzt sind (= wenige Monate), d.h. sie ist an eine prospektive Datensammlungsstrategie gebunden. An diesem Punkt gehen die Vorstellungen der Biographik und der Einzelfallanalyse auseinander. Es ist jedoch denkbar - und erste Beispiele sprechen dafür - die Einzelfallanalyse für die biographische Dokumentenanalyse i. e. S. einzusetzen (Petermann, 1982). Für die nachfolgenden methodischen Überlegungen ist es von völlig untergeordneter Bedeutung, ob die Einzelfallanalyse in der Klinischen Psychologie, Entwicklungspsychologie (Ashton, 1975; Risley & Baer, 1979), Sozialpsychologie (Feger, 1975) oder der pädagogisch-psychologischen Forschung (Kratochwill, 1977) eingesetzt wird. So ist es z.B. gut möglich, die Wachstumskurven, das subjektive Erleben von Belastungssituationen in bestimmten Lebensabschnitten, Belastung durch kritische Lebensereignisse, Variabilität im Stimmungshaushalt u.ä. zu betrachten.

Die Einzelfallanalyse unterstellt, daß man die wesentlichen Erkenntnisse über eine Einzelperson über ihre wiederholte (prospektive) Betrachtung erzielen kann. Diese Annahme ist mit bestimmten Forderungen bezüglich der Datensammlung bzw. der Versuchsplanung verbunden, die im nächsten Abschnitt referiert werden; zunächst sollen jedoch nochmals die Intentionen der Einzelfallanalyse präzisiert werden.

Die Intentionen der Einzelfallanalyse lassen sich einmal i.S. der Betrachtung der einem Merkmal innewohnenden Variabilität (a) und einmal i.S. der potentiellen Veränderbarkeit eines Merkmales (b) interpretieren. Zur Unterscheidung soll in Fall (a) von einer deskriptiven und in Fall (b) von einer explikativen Einzelfallanalyse gesprochen werden. Diese Unterscheidung zeigt auch, daß die Einzelfallanalyse nicht zwingend an das operante Paradigma gekoppelt ist (vgl. Risley & Baer, 1979), da in Fall (a) die Möglichkeit gegeben ist, ohne das Vorliegen einer klar definierten Intervention Veränderungen einzelfallanalytisch zu beschreiben. In Fall (a) ließe sich der Entwicklungsverlauf abbilden und prüfen, ob grundlegende Veränderungen im Entwicklungstrend zu ver-

zeichnen sind. Weiterhin könnte man für unterschiedliche Lebensabschnitte die der menschlichen Entwicklung innewohnende Variabilität bestimmen und charakteristische Verläufe benennen. Die so erzielte differenzierte Beschreibung des entwicklungspsychologischen Prozesses oder Krankheitsprozesses besitzt allerdings keinen explikativen Wert.

Explikative Einzelfallanalysen bestimmen den Effekt einer oder mehrerer unabhängiger Variablen auf den zeitlichen Verlauf einer Merkmalsausprägung. Der Ausdruck „unabhängige Variable“ ist in diesem Kontext weitgehend identisch mit dem Begriff „Intervention“ oder „interventionsanaloges Ereignis“, wobei durch diese Terminologie nicht zwingend eine Festschreibung des Anwendungsbereiches der explikativen Einzelfallanalyse auf den klinisch-psychologischen Bereich erfolgt. So könnte etwa für die Entwicklungspsychologie der Interventionsbegriff zunächst zweierlei bedeuten:

- (a) Die Analyse von mikrostrukturellen Prozessen innerhalb der Effektprüfung zeitlich begrenzter Interventionsmaßnahmen (z.B. zur kognitiven Frühförderung) und
- (b) die Analyse von makrostrukturellen Prozessen (= Verlaufsbeobachtung über die Lebensspanne), die von einem weitgefaßten Interventionsbegriff ausgeht; ein Beispiel hierfür stellt die Analyse von kritischen Lebensereignissen dar.

13. Einzelfallanalytische Datensammlung und Versuchsplanung

Die Durchführung von explikativen Einzelfallanalysen setzt bestimmte Schritte der Datensammlung bzw. Versuchsplanung voraus. Die Einzelfallanalyse begreift sich als prospektive Registrierung von Veränderungen, wobei eine maximale Anzahl von Daten i. S. von wiederholten Messungen pro Einzelfall gewonnen werden soll. Innerhalb der Klassifikation von klassischen Versuchsplänen (vgl. Campbell & Stanley, 1970) wird der Einzelfallversuchsplan unter dem Stichwort „Zeitreihenversuchsplan“ abgehandelt.

Bei der Gestaltung der Einzelfallversuchspläne wird über die Anzahl der wiederholten Messungen keine Aussage gemacht. Im Normalfall wird von einer großen Anzahl von Erhebungen ausgegangen (ca. 30 - 60), um eine (statistische) Aussage über die Beschaffenheit und die Variabilität des zu untersuchenden Merkmals aufstellen zu können. Die umfassendste Erfahrung im Umgang mit Einzelfallversuchsplänen liegt in der verhaltenstherapeutischen Forschung vor (vgl. Barlow & Hersen, 1977; Hersen & Barlow, 1976; Leitenberg, 1977).

Zur Illustration des Vorgehens kann auf den einfachsten Versuchsplan, das sogenannte A-B-Design, etwas genauer eingegangen werden. Bei einem A-B-Design bezeichnet A die Phase der Nicht-Behandlung und B die der Behand-

lung, wobei in beiden Phasen eine relativ große Anzahl von wiederholten Messungen erhoben werden, um das Kriteriumsverhalten, das verändert werden soll, abbilden zu können. Die erzielten Ergebnisse können in einem Koordinatensystem mit den Achsen „Zeit“ (in Minuten, Tagen etc.) und „Häufigkeiten des erwünschten Verhaltens“ abgetragen werden. Im allgemeinen wird gefordert, daß die Längen der Phasen vergleichbar sein sollen (Barlow & Hersen, 1977).

Zur Verdeutlichung des A-B-Versuchsplanes kann man von folgender Untersuchung ausgehen. Im Rahmen eines Modifikationsprogrammes soll aggressives Störverhalten abgebaut werden. Die einfachste Form der Realisierung der Studie besteht darin, nach einer Phase der Erfassung des Störverhaltens (A-Phase) eine Behandlung durchzuführen (B-Phase). Abbildung 2 stellt mögliche Ergebnisse der Baselineerhebung und des Trainingsverlaufes vor.

Es ist nun leicht vorstellbar, durch beliebige Kombinationen der A- und B-Phasen komplexe Versuchspläne zu entwickeln, mit denen eine detaillierte Analyse der unterschiedlichen Effekte möglich wird. Ein Beispiel für einen etwas komplexeren Plan bietet der Umkehrplan (A-B-A-B-Plan; vgl. Barlow & Hersen 1977).

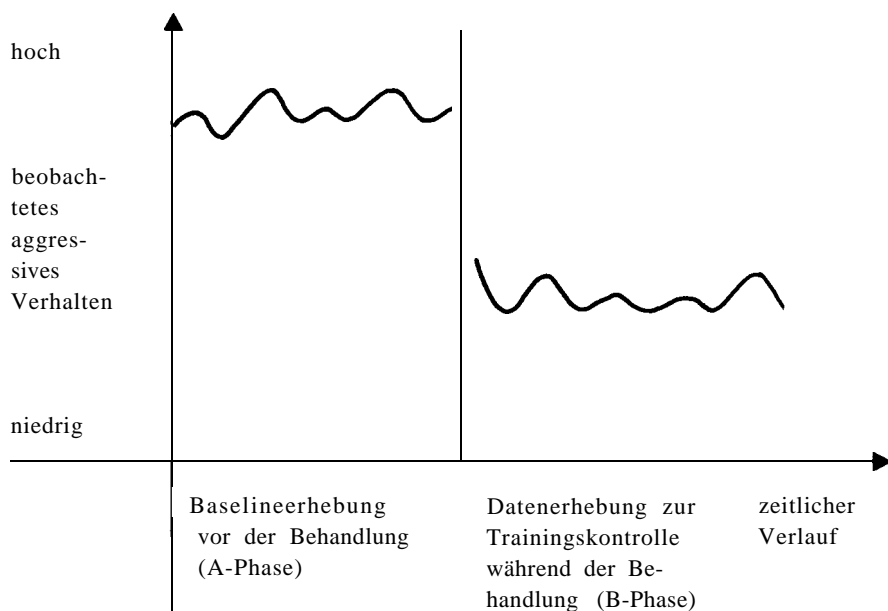


Abb. 2: Verdeutlichung des A-B-Versuchsplanes anhand fiktiver Daten.

Unter dem Aspekt der Versuchsplanung ist es notwendig, die Replikation (Wiederholung) von Einzelfallanalysen von vornherein miteinzubeziehen. Solche Bemühungen lassen sich einerseits als direkte und andererseits als systematische Replikation in Angriff nehmen.

(a) *Direkte Replikation:* Die Wiederholung des Vorgehens bei derselben Person (= intraindividuelle Replikation) und die Wiederholung mit anderen Personen (= interindividuelle Replikation) muß unterschieden werden. Eine intrasubjektive Replikation ist an die Beschaffenheit der Zeitbereichsvariablen gebunden, d.h. eine Replikation anhand einer Person ist nur möglich, wenn sich über den Untersuchungszeitraum zwischen der ersten und zweiten (Wiederholungs-) Messung keine Veränderung vollzieht, die die Vergleichbarkeit der Ergebnisse einschränkt. Eine intersubjektive Replikation setzt die Vergleichbarkeit der untersuchten Personen hinsichtlich der entscheidenden Variablen voraus.

(b) *Systematische Replikation:* Unter der systematischen Replikation verstehen Hersen & Barlow (1976) den Versuch, die Ergebnisse einer Serie von direkten Replikationen bei unterschiedlichen Versuchsbedingungen, Versuchsleitern, Verhaltensweisen u.a. zu wiederholen. Damit strebt die systematische Replikation im Gegensatz zur direkten Replikation eine Verallgemeinerung mit Hilfe einer Vielzahl von Einzelfallanalysen an, die durch die systematische Variation bestimmter, den Aussagegehalt potentiell moderierender Variablen (situative Aspekte der Versuchsplanung, Instruktion des Experimentes, Versuchsleitervariablen) den Gültigkeitsbereich der Aussage evaluiert.

Die bisherigen Überlegungen zur Replikation von Einzelfallanalysen machen deutlich, daß beim einzelfallanalytischen Vorgehen die Akribie, die bei der Planung und Durchführung dieser Studien an den Tag gelegt werden muß, oft ein Vielfaches von dem sein sollte, das man traditionellerweise beim klassischen Experiment für nötig erachtet.

14. *Übersicht über statistische Auswertungsmethoden für Einzelfälle*

Die große Anzahl der wiederholten Messungen bei Einzelfallbetrachtungen hat dazu geführt, daß sich eine Reihe von Auswertungsmethoden entwickelte, die die Effekte aus wiederholten Messungen (= seriale Abhängigkeit) berücksichtigen. Unter serialer Abhängigkeit versteht man die Stärke der Beeinflussungstendenz, die sich in dem Ausmaß des „Erinnerns“ eines Meßwertes an vorhergegangene niederschlägt. Dieses „Erinnern“ führt bei der Anwendung der herkömmlichen Auswertungsverfahren dazu, daß Veränderungen aufgezeigt werden, denen kein realer Prozeß zugrundeliegt und die somit als Artefakte zu interpretieren sind. Es ist nun zu fragen, wie man die herkömmlichen

statistischen Auswertungsmodelle modifizieren muß, damit es mit ihnen gelingt, die seriale Abhängigkeit von der realen Veränderung zu trennen.

Zur Lösung der angesprochenen Fragen wird eine Reihe von Vorschlägen unterbreitet, von denen einige wichtige in Abbildung 3 enthalten sind. Die Übersicht gliedert sich, nach dem den Daten zugrundegelegten Skalenniveau, in qualitative (beim Vorliegen von Nominal- und Ordinalskalenniveau) und quantitative Ansätze (beim Vorliegen von Intervallskalenniveau). Die Einordnung der Auswertungsmethoden in die Abbildung erfolgt auf der Grundlage der minimal notwendigen Skalenniveaus. Für die unterschiedlichen Skalenniveaus lassen sich zudem die anfallenden Informationen graphisch aufbereiten (vgl. Parsonson & Baer, 1978; Revenstorf & Vogel, 1979).

Einen guten Überblick über Auswertungsmethoden für qualitative Daten geben Gottman & Notarius (1978) und Revenstorf & Vogel (1979), die u.a. Markoffanalysen erster und höherer Ordnung sowie informationstheoretische Betrachtungsansätze vorstellen. Weitere Auswertungsmethoden, die nicht in Abbildung 3 enthalten sind, stellen Huber (1978) und Petermann (1977) zusammen.

Für die Analyse quantitativer Daten steht zudem eine große Auswahl von Verfahren zur Verfügung, die zumindest die folgenden Ansätze beinhaltet:

- die varianzanalytische Auswertung (z.B. die Shine-Bower-Analyse und das Vorgehen von Gentile et al.; vgl. hierzu Petermann, 1978),
- die faktorenanalytische Auswertung (P- und O-Technik; Cattell, 1977; mit Einschränkungen die drei-modale Faktorenanalyse nach Tucker) und
- die zeitreihenanalytische Auswertung i.e. S. (vgl. deterministische Ansätze: Beschreibung durch Polynome und die sogenannte „Fourier-Analyse“; vgl. Revenstorf & Keeser, 1979; stochastische Ansätze: Beschreibung durch autoregressive Prozesse und Gleitmittelprozesse; vgl. Dahme, 1977; Gottman & Glass, 1978; Keeser, 1980; Revenstorf & Keeser, 1979).

Neben den in Abbildung 3 aufgeführten statistischen Auswertungsmethoden werden vor allem in der klinisch-psychologischen Forschung Verfahren zur „qualitativen“ Einzelfallanalyse diskutiert. Qualitative Einzelfallanalysen möchten durch Befragung (z.B. durch Intensiv-Interviews) den subjektiven Problemraum des Patienten erfassen, um so eine bessere Berücksichtigung dieser Aspekte im therapeutischen Prozeß zu gewährleisten (vgl. Bannister, 1977; Kiresuk & Sherman, 1968; Rüppell, 1979). Es ist zu vermuten, daß durch diesen Ansatz einer qualitativen Einzelfallanalyse, der sicherlich in der klinisch-psychologischen Forschung (Psychotherapieforschung) noch an Bedeutung gewinnen wird, die ursprünglichen Intentionen der biographischen Methode noch stärker ihren Niederschlag finden werden.

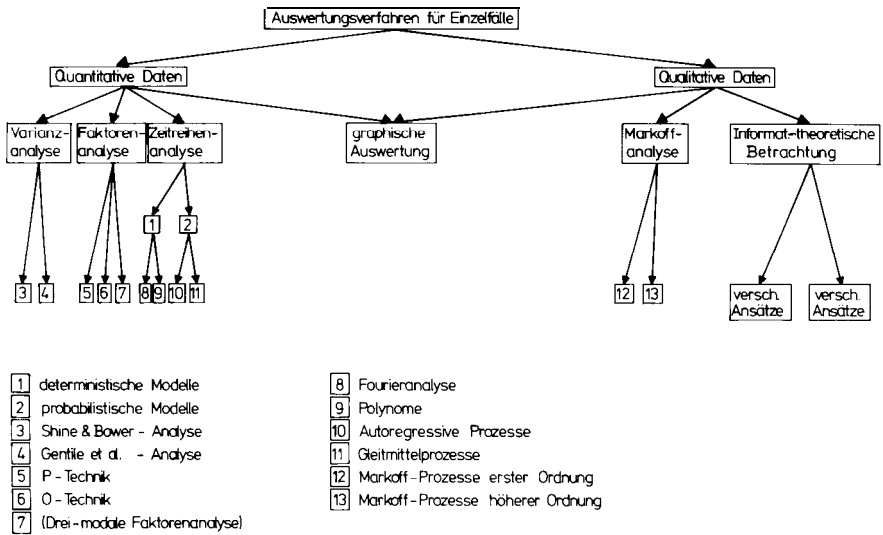


Abb. 3: Übersicht über mögliche Auswertungsverfahren für Einzelfallanalysen (entnommen aus: Petermann, 1981, S. 40).

Literatur

- Allport, G. W. 1937. *Personality. A psychological interpretation.* New York: Holt.
- Allport, G. W. 1942. *The use of personal documents in psychological science.* New York: New York Academy at Science.
- Allport, G. W. 1969. *Pattern and growth in personality.* New York: Holt, Rinehart & Winston, 1962. Dt. Ausgabe, hrsg. von H. v. Bracken unter dem Titel: *Gestalt und Wachstum der Persönlichkeit.* Meisenheim: Hain.
- Allport, G. W. 1965. (Ed.) *Letters from Jenny.* New York: Harcourt, Brace & World.
- Anton, J. L. 1978. *Studying individual change.* In L. G. Goldman (Ed.), *Research and the counselor.* New York: Wiley.
- Ashton, P. T. 1975. *Cross-cultural Piagetian research: An experimental perspective.* *Harvard Educational Review*, 4, 475-506.
- Bahrdt, H. P. 1975. *Erzählte Lebensgeschichten von Arbeitern.* In M. Osterland (Hrsg.), *Arbeitssituation, Lebenslage und Konfliktpotential.* Frankfurt: Sozietätsverlag.
- Bannister, D. 1977. (Ed.) *New perspectives in personal construct theory.* London: Academic Press.
- Bain, R. K. 1960. *The researcher's role.* In R. N. Adams & J. J. Preiss (Eds), *Human Organisation research.* Homewood, Ill.: Dorsey Press 1960, 146-152.
- Baer, D. M., Wolf, M. M. & Risley, T. R. 1968. *Some current dimensions of applied behavior analysis.* *Journal of Applied Behavior Analysis*, 1, 91-97.
- Barlow, D. H. & Hersen, M. 1977. *Designs für Einzelfallexperimente.* In F. Petermann (Hrsg.), *Methodische Grundlagen Klinischer Psychologie.* Weinheim: Beltz.
- Barker, R. G. & Wright, H. F. 1955. *Midwest and its children.* New York: Harper.
- Bartenwerfer, H. & Giesen, H. 1973. *Pilotstudie über die Beobachtung und Analysen von Bildungslebensläufen.* Frankfurt: Deutsches Institut für Internationale Pädagogische Forschung.
- Bellak, I. 1964. *Somerset and Maugham: a thematic analysis of ten short stories.* In R. W. White (Ed.), *The study of lives.* New York: Atherton.
- Binswanger, L. 1928. *Lebensfunktion und innere Lebensgeschichte.* *Monatsschrift für Psychiatrie*, 68, 52-129.
- Binswanger, L. 1942. *Grundformen und Erkenntnis menschlichen Daseins.* Zürich: Rascher.
- Bühler, C. 1927. *Tagebuch eines jungen Mädchens. Quellen zum Studium der Jugendkunde.* Jena: Fischer, 1922; 2. Auflage unter dem Titel: *Tagebücher von zwei Mädchen.*
- Bühler, C. 1933. *Der menschliche Lebenslauf als psychologisches Problem.* Leipzig: Hirzel.
- Bühler, C. & Massarik, F. 1969. (Hrsg.) *Lebenslauf und Lebensziele.* Stuttgart: Fischer.

- Burgess, E. W. 1945. Research methods in sociology. In E. Gurvitch & G. Moore (Ed.), *Twentieth century sociology*. New York: Columbia University Press.
- Campbell, D. T. & Stanley, J. C. 1970. Experimentelle und quasi-experimentelle Anordnungen in der Unterrichtsforschung. In K. H. Ingenkamp (Hrsg.), *Handbuch der Unterrichtsforschung*. Weinheim: Beltz.
- Cattell, R. B. 1977. Die Erfassung von Veränderungen mit der P-Technik und der inkrementellen R-Technik. In F. Petermann (Hrsg.), *Methodische Grundlagen Klinischer Psychologie*. Weinheim: Beltz.
- Champeys, F. H. 1881. Notes on an infant. *Mind*, 14, 104-107.
- Dahme, B. 1977. Zeitreihenanalyse und psychotherapeutischer Prozeß. In F. Petermann (Hrsg.), *Methodische Grundlagen Klinischer Psychologie*. Weinheim: Beltz.
- Dailey, C. A. 1972. The uses of autobiography in coping with problems of identity formation among undergraduates. Sudburg: Mass. Biodata, Inc.
- Dailey, C. A. 1975. The future of biographical psychology. Unveröffentlichtes Manuskript.
- Darwin, C. 1877. A biographical sketch of an infant. *Mind*, 2, 285-294.
- Davis, A. & Dollard, J. 1940. Children of bondage: the personality development of negro youth in the urban south. Washington: American Council of Leducation.
- DeBoor, C. 1966. Hysterie: Konversionsneurotisches Symptom oder Charakterstruktur? *Psyche*, 20, 588-599.
- DeWaele, J. P. 1974. The role of autobiography in personality assessment. Paper presented at the International Congress of Applied Psychology, Montreal.
- Dilthey, W. 1924. Ideen über eine beschreibende und zergliedernde Psychologie (1894). In Dilthey, W. (Hrsg.), *Gesammelte Schriften*. Band V. Leipzig: Teubner.
- Dollard, J. 1935. *Criteria for the life history*. New Haven: Yale University Press.
- Dukes, W. F. 1977. N = 1. In F. Petermann (Hrsg.), *Methodische Grundlagen Klinischer Psychologie*. Weinheim: Beltz.
- Edinger, L. J. 1964. Political Science and political biography: reflections on the study of leadership. *Journal of Politics*, 36, 34-50.
- Erikson, H. E. 1950. Growth and crises of the healthy personality. In M. Senn (Ed.), *Symposion on the healthy personality*. New York: Macey Foundation.
- Erikson, H. E. 1968. On the nature of psychohistorical evidence: in search of Ghandi. *Daedalus*, 97, 695-730.
- Feger, H. 1975. Längsschnittliche Erfassung intraindividuelle Unterschiede bei Einstellungsstrukturen. In U. Lehr & F. Weinert (Hrsg.), *Entwicklung und Persönlichkeit*. Stuttgart: Kohlhammer.
- Freud, S. 1940-1950. *Gesammelte Werke*. Bd. I - XVII. London: Imago.
- Friedrichs, J. & Kamp, K. 1978. Methodologische Probleme des Konzeptes „Lebenszyklus“. In M. Kohli (Hrsg.), *Soziologie des Lebenslaufs*. Darmstadt: Luchterhand.

- Garrison, K. R. & Muchinsky, P. M. 1977. Attitudinal and biographical predictors of incidental absenteeism. *Journal of Vocational Behavior*, 10, 221-230.
- Gesemann, G. 1924. Grundlagen einer Charakterologie Gogols. In *Jahrbuch für Charakterologie*. Bd. 1. Berlin: Mitter.
- Glad, B. 1973. Contributions of psychobiography. In J. N. Knutson (Ed.), *Handbook of Political Psychology*. San Francisco: Jossey Bass.
- Goldstein, H. 1979. The design and analysis of longitudinal studies. London: Academic Press.
- Gottman, J. M. & Notarius, C. 1978. Sequential analysis of observational data using Markov chains. In T. R. Kratochwill (Ed.), *Strategies for evaluating change*. New York: Academic Press.
- Gruhle, H. W. 1952. *Geschichtsschreibung und Psychologie*. Bonn: Bouvier.
- Hambitzer, M. 1962. Schicksalsbewältigung und Daseinsermöglichung bei Körperbehinderten. Bonn: Bouvier.
- Herder, J. G. v. 1778. *Vom Erkennen und Empfinden der menschlichen Seele*. Leipzig: Hartknoch.
- Hersen, M. & Barlow, D. H. 1976. Single-case experimental designs: Strategies for studying behavior change. New York: Pergamon.
- Holtzman, W. H. 1977. Statistische Modelle zur Untersuchung von Veränderungen im Einzelfall. In F. Petermann (Hrsg.), *Methodische Grundlagen Klinischer Psychologie*. Weinheim: Beltz.
- Horner, A. J. 1969. Die Entwicklung von Zielen im Leben von Charence Darrow. In C. Bühler & F. Massarik (Hrsg.), *Lebenslauf und Lebensziele*. Stuttgart: Fischer.
- Huber, H. P. 1978. Kontrollierte Fallstudie. In L. J. Pongratz (Hrsg.), *Handbuch der Psychologie*. Band 7.2. Göttingen: Hogrefe.
- Jaehner, D. 1930. *Zwei Tage aus dem Leben dreier Geschwister*. Leipzig: Barth.
- Jaspers, K. 1962. *Allgemeine Psychopathologie*. Berlin: Springer, 1913; 6. Auflage.
- Jones, E. 1960-1962. *Das Leben und Werk von Sigmund Freud*. 3 Bd. Bern: Huber.
- Kardiner, A. 1945. *Psychological frontiers of Society*. New York: Columbia University Press.
- Kazdin, A. E. 1978. Methodological and interpretative problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology*, 46, 629-642.
- Keeser, W. 1980. *Zeitreihenanalyse in der Klinischen Psychologie*. Ein empirischer Beitrag zur Box-Jenkins Methodologie. München: Phil. Diss.
- Kessen, W., Haith, M. M. & Salapatek, P. H. 1970. Human infancy: a bibliography and guide. In P. H. Mussen (Ed.), *Carmichaels Manual of Child Psychology*. New York: Wiley.
- Kinsey, A. C. Pomeroy, W. B., Martin, C. E. & Gebhard, P. H. 1953. *Sexual behavior of human female*. Philadelphia, (dt. Ausgabe. Frankfurt: Fischer, 1963).

- Kipnowski, A. 1980. Formen der Daseinsbewältigung bei chronischer Krankheit. Eine Untersuchung an erwachsenen Hämophilen. Bonn: Phil. Diss.
- Kirchner, F. Th., Kissel, E., Petermann, F. & Böttger, P. 1977. Interne und Externe Validität empirischer Untersuchungen in der Psychotherapieforschung. In F. Petermann (Hrsg.), Psychotherapieforschung. Weinheim: Beltz.
- Kiresuk, T. J. & Sherman, R. E. 1968. Goal attainment scaling: a general method for evaluating comprehensive community mental health program. *Community Mental Health Journal*, 4, 443-453.
- Kitay, Ph. M. 1963. Symposium on reinterpretations of the Schreber case: Freud's theory of Paranoia. I. Introduction. *International Journal of Psychoanalysis*, 4, 191-194.
- Kohli, M. 1977. Lebenslauf und Lebensmitte. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 29, 625-656.
- Kohli, M. 1978. (Hrsg.), *Soziologie des Lebenslaufs*. Darmstadt: Luchterhand.
- Kratochwill, T. R. 1977. N = 1. An alternative research strategy for school psychologists. *Journal of School Psychology*, 15, 239-249.
- Kratochwill, T. R. 1978. (Ed.) *Single subject research. Strategies for evaluating change*. New York: Academic Press.
- Kuiper, P. C. 1965. Dilthey's Psychologie und ihre Beziehung zur Psychoanalyse. *Psyche*, 19, 241-249.
- Kuiper, P. C. 1966. Die Psychoanalyse der schöpferischen Persönlichkeit. *Psyche*, 20, 104-127.
- Lampl de Groot, J. 1963. Symptombildung und Charakterbildung. *Psyche*, 17, 1-21.
- Leitenberg, H. 1977. Einzelfallmethodologie in der Psychotherapieforschung. In F. Petermann & C. Schmook (Hrsg.), *Grundlagentexte der Klinischen Psychologie*. Bd. 1. Bern: Huber.
- Lehr, U. 1969. *Frau im Beruf. Eine psychologische Analyse der weiblichen Berufsrolle*. Frankfurt: Athenäum.
- Lehr, U. 1978. Kontinuität und Diskontinuität im Lebenslauf. In L. Rosenmayr (Hrsg.), *Die menschlichen Lebensalter*. München: Piper.
- Lehr, U. & Thomae, H. 1958. Eine Längsschnittuntersuchung bei menschlichen Angestellten. *Vita humana*, 1, 100-110.
- Lehr, U. & Thomae, H. 1965. *Konflikt, seelische Belastung und Lebensalter*. Opladen: Westdeutscher Verlag.
- Lewis, O. 1961. *The children of Sanchez. Autobiography of a Mexican family*. New York: Random House.
- Linton, R. 1945. *The cultural background of personality*. New York: Appleton Century Press.
- Loch, W. 1966. über einige Strukturmerkmale und Funktionen psychoanalytischer Deutungen. *Psyche*, 20, 377-396.

- Loewenstein, R. M. 1960. Variationen der psychoanalytischen Technik. *Psyche*, 13, 594-608.
- Maslow, A. 1950. *Self-actualizing people: a study of psychological health*. New York: Grune & Stratton.
- Mazlish, B. 1972. *In search of Nixon: a psychohistorical inquiry*. New York: Basic Books.
- McGuire, W. J. 1973. The Yin and Yang of progress in social psychology: seven koan. *Journal of Personality and Social Psychology*, 26, 446-456.
- Mierke, K. 1944. Psychologische Diagnostik. In N. Ach (Hrsg.), *Lehrbuch der Psychologie*. Bamberg: Buchner.
- Miettinen, O. S. 1970. Matching and design efficiency in retrospective studies. *American Journal of Epidemiology*, 91, 111-118.
- Misch, G. 1907-1962. *Geschichte der Autobiographie*. 3 Bd. Frankfurt: Klostermann.
- Neufeld, R. W. J. 1977. *Clinical quantitative methods*. New York: Grune & Stratton.
- Niederland, W. G. 1963. Further data and memorabilia pertaining to the Schreber case. *International Journal of Psychoanalysis*, 44, 201-207.
- Obst, G. H. 1961. Eine Untersuchung an Werkmeistern über Einflüsse auf ihre Persönlichkeitsentwicklung. Erlangen: Phil. Diss.
- Olbrich, M. 1967. *Formale Analyse der Verhaltensstruktur auf Grund des Tageslaufs*. Bonn: Psychologische Vorexamensarbeit.
- Osterland, M. 1973. Lebensgeschichtliche Erfahrung und gesellschaftliches Bewußtsein. Anmerkungen zu einer soziobiographischen Methode. *Soziale Welt*, 84, 409-417.
- Osterland, M. 1978. Lebensbilanzen und Lebensperspektiven von Industriearbeitern. In M. Kohli (Hrsg.), *Soziologie des Lebenslaufes*. Neuwied: Luchterhand.
- Parsonson, B. S. & Baer, D. M. 1978. The analysis and presentation of graphic data. In T. R. Kratochwill (Ed.), *Single subject research. Strategies for evaluating change*. New York: Academic Press.
- Paul, S. 1979. *Begegnungen für geschichtspersönliche Dokumente*. LB 1. Hohen-schaftlarn: Reuer.
- Pauleinkhoff, B. 1963. Die Rolle des Tageslaufes in der Persönlichkeits- und Ganzheitspsychologie. *Archiv für gesamte Psychologie*, 117, 67-77.
- Petermann, F. 1977. (Hrsg.) *Methodische Grundlagen Klinischer Psychologie*. Weinheim: Beltz.
- Petermann, F. 1978. *Veränderungsmessung*. Stuttgart: Kohlhammer.
- Petermann, F. 1980. Erfassung und quantitative Beschreibung von Einstellungsänderungen. In F. Petermann (Hrsg.), *Einstellungsmessung - Einstellungsforschung*. Göttingen: Hogrefe.
- Petermann, F. 1981. Möglichkeiten der Einzelfallanalyse in der Psychologie. *Psychologische Rundschau*, 32, 31-48.

- Petermann, F. 1982. Einzelfalldiagnose und klinische Praxis. Stuttgart: Kohlhammer.
- Petermann, F. & Petermann, U. 1980. Erfassungsbogen für aggressives Verhalten in konkreten Situationen (EAS-J; EAS-M). Braunschweig: Westermann.
- Preyer, W. 1882. Die Seele des Kindes. Leipzig: Grieben.
- Rasmussen, V. 1931. Diary of a child's life from birth to the fifteenth year. London: Glydendal.
- Rasmussen, V. 1934. Ruth: Tagebuch über die Entwicklung eines Mädchen von der Geburt bis zum 18. Lebensjahr. Berlin: Oldenbourg.
- Revenstorf, D. & Keeser, W. 1979. Zeitreihenanalyse von Therapieverläufen - ein Überblick. In F. Petermann & F.-J. Hehl (Hrsg.), Einzelfallanalyse. München: Urban & Schwarzenberg.
- Revenstorf, D. & Vogel, B. 1979. Zur Analyse qualitativer Verlaufsdaten - ein Überblick. In F. Petermann & F.-J. Hehl (Hrsg.), Einzelfallanalyse. München: Urban & Schwarzenberg.
- Richardson, S. A., Dohrenwend, B. S. & Klein, D. 1965. Interviewing. Its forms and functions. New York: Basic Books.
- Risley, T. R. & Wolf, M. M. 1979. Strategien zur Untersuchung von Veränderungen des Verhaltens in der Zeit. In P. B. Baltes (Hrsg.), Entwicklungspsychologie der Lebensspanne. Stuttgart: Klett-Cotta.
- Robinson, P. W. & Foster, D. F. 1979. Experimental Psychology: a small-N approach. New York: Harper & Row.
- Romein, J. Die Biographie. Bern (o.J.)
- Rüppell, M. 1979. Einzelfallanalyse und Subjekt. In F. Petermann & F.-J. Hehl (Hrsg.) Einzelfallanalyse. München: Urban & Schwarzenberg.
- Schmidt, H. D. 1977. „Der Unglücklichste aller Sterblichen“. In L. Goldschmidt (Hrsg.) Um die unsterbliche Geliebte. Eine Bestandsaufnahme. Beethoven-Studien 2. Leipzig: VEB Deutscher Verlag für Musik.
- Schraml, W. J. 1965. Die Psychoanalyse und der menschliche Lebenslauf. Psyche, 19, 250-268.
- Scharnweber, 1979. Die Auseinandersetzung von Dialysepatienten mit ihrer Lebenssituation. Bonn: Psychologische Diplomarbeit.
- Scupin, E. & Scupin, G. 1907. Bubi im ersten bis dritten Lebensjahre. Leipzig: Grieben.
- Sears, R. R. 1974. Episodic and content analysis of Mark Twain's novels: Implications for autobiography. Paper presented at Symposion ‚Psychology in Biography Autobiography‘. 18th International Congress of Applied Psychology, Montreal.
- Seligman, E. 1979. Erlernte Hilflosigkeit. München: Urban & Schwarzenberg.
- Shinn, M. W. 1900. The biography of a baby. New York: Houghton Mifflin.
- Simonton, D. K. 1975. Sociocultural context of individual creativity: a transhistorical time-series analysis. Journal of Personality and Social Psychology, 32, 1119-1133.

- Simonton, D. K. 1976. Biographical determinants of achieved eminence: a multivariate approach to the Cox data. *Journal of Personality and Social Psychology*, 33, 318-326.
- Simonton, D. K. 1977a. Cross-sectional time-series experiments: some suggested statistical analysis. *Psychological Bulletin*, 84, 489-502.
- Simonton, D. K. 1977b. Creative productivity, age, and stress: a biographical time-series analysis of 10 classical composers. *Journal of Personality and Social Psychology*, 35, 791-804.
- Spranger, E. 1963. *Menschenleben und Menschenfragen*. München: Piper.
- Spranger, E. 1966. *Psychologie des Jugendalters*. Heidelberg: Quelle & Meyer; 20. Auflage.
- Stern, W. & Stern, C. 1928/31. *Monographien über die seelische Entwicklung des Kindes*. Leipzig: Barth; 4. Auflage in 2 Bd.
- Swaminathan, H. & Algina, J. 1977. Analysis of quasi-experimental time-series designs. *Multivariate Behavioral Research*, 12, 111-131.
- Tack, W. H. 1979. Testtheoretische Grundlagen der Einzelfallanalyse. In F. Petermann F.-J. Hehl (Hrsg.) *Einzelfallanalyse*. München: Urban & Schwarzenberg.
- Taine, H. M. 1877. Taine on the acquisition of language by children. *Mind*, 2, 252-259.
- Thomae, H. 1968. *Das Individuum und seine Welt*. Göttingen: Hogrefe.
- Thomae, H. & Kranzhoft, H. E. 1979. Erlebte Unveränderlichkeit von gesundheitlicher und ökonomischer Belastung. *Zeitschrift für Gerontologie*, 12, 439-459.
- Thomas, W. I. & Znaniecki, F. 1918-1920. *The polish peasant in Europe and America*. 5 vols. Boston: Badger.
- Tiedemann, D. 1787. Beobachtungen über die Entwicklung der Seelenfähigkeiten bei Kindern. *Hessische Beiträge zur Gelehrten Kunst*, 2, 313-333; 3, 486-502.
- White, R. W. 1964. (Ed.) *The study of lives. Essays on personality in honor of Henry A. Murray*. New York: Atherton Press.
- Whyte, W. F. 1955. *Street corner society*. Rev. ed. Chicago: University of Chicago Press.
- Wieser, I. 1973. Biographische Methode. In C. Müller (Hrsg.), *Lexikon der Psychiatrie*. Berlin: Springer.
- Wilson, R. N. 1964. Albert Camus: Personality as creative struggle. In R. W. White (Ed.), *The study of lives*. New York: Atherton.
- Wolfenstein, E. v. 1967. *The revolutionary personality: Lewin, Trotsky, and Gandhi*. Princeton: Princeton University Press.

Sach-Register

Anamnese 322

Antworttendenzen 230, 234, 260f

Bedeutung 164

affektive - 164

konnotative - 164

denotative - 164

Befragung 302ff, 321f

Begriffsbestimmung 301f

Fehlerquellen bei - 306, 310f

-, Methodenvarianten 303f

mündliche - 308f

Realkontakt - 313f

schriftliche - 305f

telefonische - 314f

- von Kindern 315f

Behaviorismus 89ff

Beobachtung(en) 1ff

-, Auswahl des zu beobachtenden
Verhaltens 18ff

-, Auswahl und Schulung von
Beobachtern 17f

Auswahl von Untersuchungs-
personen bei - 16ff

Beobachtungseinheit 11ff

Bewertung von - 22ff

-, Kategoriensystem 12ff

Planung von - 6ff

Reliabilität von - 22ff

Reproduzierbarkeit von - 22f

teilnehmende - 4, 5f

Validität von - 22f, 48ff

wissenschaftliche(n) - 1ff

Arten von - 3ff

Bias (Interviewer-Bias) 330

Biographische(n) Methode

(Biographik) 362ff

- in der Entwicklungs-
psychologie 365f

Forderungen an - 376

- in der humanistischen
Psychologie 367f

Idiographische Persön-
lichkeitspsychologie und - 364f

Objektivität der - 375ff

-, Probleme in der
Psychoanalyse 368ff

psychologische Streßforschung und
- 373ff

- in der Sozialisationsforschung 371ff
statistische Auswertung 386f

Butler & Haigh-Q-Sort 135ff

California Q-Set (CQ-Set) 136

Denken, lautes 91

Einzelfallanalyse 387ff

-, Datensammlung und Ver-
suchsplanung 389ff

-, statistische Auswer-
tungsmethoden 391ff

EPA-Struktur 159f

Ereignisstichprobe 20f

Erkundungsgespräch s. Exploration

Erlebnisbeschreibung 77f

Experiment 9, 326

Exploration 321ff, 383

Auswertung einer - 334f

- in der Eignungsdiagnostik 349ff
Geschichte der - 323f

methodische Prinzipien der - 329ff

-, qualitative Charakterisierung 325ff

-, Reliabilität 339f

-, Validität 345f

Exploratives Gespräch s. Exploration

Explorative Methode s. Exploration

- Facettentheorie 6ff, 15
- Feldbeobachtung 9
- Frage(n) 241
 - bedeutungsäquivalente - 261f
 - heikle - 274f, 277f
 - offene, geschlossene
 - Reihenfolge 267f, 270, 273f
- Fragebogen 222ff, 327
 - , Antwortkategorien 254f
 - , Antwortmotivation 269ff
 - Antwortverhalten in - 229f, 231f
 - Anwendungsgebiete von - 227f
 - äußere Gestaltung 279ff
 - Beantwortungsfehler in - 235
 - , Fragenformulierung 251ff, 261ff
 - , Frageninhalte 243f, 252ff
 - , Fragentypen 241, 246
 - , Gedächtnisprozesse 236f
 - Konstruktionsgesichtspunkte
 - von - 224f
 - , kontexteffekte 270ff
 - Systematik von - 222f
 - Verfälschbarkeit von - 229f
- Fremdbeobachtung 77ff
 - Verlässlichkeit der - 118
- Generalisierbarkeit(stheorie) 8, 41f
- Gespräch 321
- Halo-Effekt 104
- Idiographische Persönlichkeitspsychologie 364f
- Interview 322, 327
 - halbstrukturiertes - 383
- Introspektion 78ff
- ipsative Daten 135
- Kappa (x) von Cohen 30ff
- Kategoriensystem 12f
- Korrelation(skoeffizienten) (s.a. Übereinstimmungsmaße)
 - Intraklassen - 37ff
 - Produkt-Moment - 25, 37, 38
 - Rang - 34, 36
- Metakognition 97
- multitrait-multimethod matrix (MTMM) 51f
- Personen-Stichprobe 20
- Pi (n) von Scott 30
- Psychobiographie 370
- Q-Korrelation 135
- Q-Sort(-Verfahren) 135ff
 - , Anwendung 140f
 - , Auswertung 144
 - Butler & Haigh - 136ff, 142
 - , California Q-Set (CQ-Set) 136, 147
 - , Einordnung 135
 - , Itemselektion und Itemorganisation 141f
 - , Reliabilität 144f
 - , Validität 145f
- R-Daten 147
- Reliabilität 22, 23f, 46f
 - , pfadanalytische Modelle 42ff
 - varianzanalytische Ansätze 37ff
- Reproduzierbarkeit von (wissenschaftl.) Beobachtungen 22f
- Rho (q) von Spearman 34, 36
- S-Daten 147
- Schätzskalensystem 13
- Selbstbeobachtung 77ff
 - methodische - 83ff
- Verlässlichkeit der - 118
- Semantisches Differential 154ff
 - , Auswahl der Skalen 189ff
 - , Daten-Reduktionstechniken 161ff
 - , EPA-Struktur 151f
 - , interindividuelle Unterschiede 169ff
 - , Konzept-Skaleninteraktion 172ff
 - , Konzeptvarianz 162ff
 - , Reliabilität 181ff
 - , Skalenqualität 184ff
 - , systematische Urteilsfehler 178ff

- , transkulturelle Stabilität 168f
- zugrundeliegende Modelle 154ff
- Signalerkennungstheorie 107
- Übereinstimmung(smaße) 26ff
 - für Intervalldaten 37ff
 - , pfadanalytische Modelle 42ff
 - varianzanalytische Ansätze
 - , (Intraklassenkoeffizienten) 37ff
 - für Nominaldaten 29ff
 - , Π (π) 30
 - , G-Index 33
 - Kappa (κ) 30ff
 - , für Ordinaldaten 36f
 - , Gamma (γ) 36
 - , Rho (ρ) 34, 36
 - , W von Kendall 36
- prozentuale - (%Ü) 26ff
- Validität 22, 48ff
 - Konstrukt- 50ff
 - konvergierende und diskriminierende
 - 8, 51ff
 - Kriteriums- 49
- Verfälschbarkeit (von Fragebogen) 229f
- Verhaltensbeobachtung 99
- Verhaltenseinschätzung 117
- Vorstellungen, bildhafte 98f
- W-Konkordanzmaß von Kendall 36
- Zeichensystem 13
- Zeitstichprobe 18
- Zusammenhang 29

Autoren-Register

Hinweis: Die kursivgedruckten Seitenangaben beziehen sich auf die Literaturverzeichnisse der Artikel.

- Ach, N. 84ff, 88, 124
 Adair, J. 98, 124
 Adams, F. M. 200
 Adams, J. S. 260, 291
 Adams, R. S. 14, 60
 Adler, A. 370, 378
 Aheshwari, A. 207
 Aiken, E. G. 200
 Aivano, S. L. 62
 Ajzen, I. 10, 64
 Akuto, H. 216
 Alexander, S. 199, 200
 Algina, J. 387, 400
 Allison, R. B. 200
 Althausen, R. P. 57, 60
 Allport, G. W. 364, 371, 394
 Alutto, J. A. 287, 291
 Alwin, D. F. 57, 60, 75
 Amerine, M. A. 36, 60
 Amsfeld, Elizabeth 200
 Amster, Harriet 200
 Anastasi, A. 224, 229, 291
 Andersen, E. B. 231, 291
 Anderson, A. B. 168, 171, 185, 188, 200
 Anderson, R. L. 47, 71
 Anger, H. 241, 245, 248, 251f, 253, 257, 261, 264, 269f, 291, 308, 318, 322, 325, 356
 Angst, J. 292, 307, 319
 Anisfeld, M. 200
 Anton, J. L. 388, 394
 Antons, K. 356
 Arnold, J. B. 168, 201
 Arnold, W. 356
 Armstrong, J. S. 307, 318
 Arntzen, F. 304, 318
 Arrington, R. E. 18, 61
 Arthur, A. Z. 178, 179, 201
 Asendorpf, J. 26, 32, 37, 61
 Ash, R. 240, 289, 291
 Ashley, R. K. 271, 278, 301
 Ashton, P. T. 388, 394
 Assakul, K. 47, 61
 Assessment Staff 324, 356
 Atteslander, P. 223, 241, 291, 305, 318
 Atwood, J. T. 186, 201
 Avison, W. R. 57, 61
 Azrin, N. H. 123, 124
 Baade, W. 356
 Babbie, E. R. 306, 318
 Baton, F. 83, 124
 Baer, D. M. 122, 124, 128, 387, 392, 394, 398
 Bakan, D. 90, 124
 Bahrdrdt, H. P. 372, 394
 Bailey, K. D. 305, 310, 315, 318
 Bain, R. K. 332, 356
 Baker, F. B. 45, 54, 67
 Bales, R. F. 11, 13, 61
 Ballweg, J. A. 173, 177, 210
 Baltes, P. B. 305, 318
 Bannister, D. 392, 394
 Barclay, A. 201
 Barker, R. G. 377, 394
 Barkus, Ph. 334, 359
 Barker, R. 20, 62, 100, 124
 Barlow, D. H. 148, 151, 364, 388, 389, 394, 396
 Barnard, J. W. 201
 Barnard, P. 263, 281, 301
 Barnes, J. 286f, 296
 Barton, A. H. 245, 247, 291, 297
 Barrett, G. V. 201
 Bartenwerfer, H. 372, 394
 Bartko, J. J. 37, 61
 Bass, A. R. 210
 Baumann, D. J. 144, 148
 Baumann, U. 199, 201
 Baumgartner, R. 307, 319
 Baxter, J. C. 201
 Bean, A. G. 212
 Beck, R. C. 202
 Beck, W. 324, 356
 Behrens, K. C. 291
 Bellak, I. 368, 394
 Bellamy, G. T. 125, 128
 Bellows, R. 324, 356
 Belschner, W. 317, 318
 Belson, W. A. 256, 291
 Bern, D.J. 93, 125
 Benel, Denise, C. R. 201
 Benel, R. A. 201
 Bennett, L. F. 151
 Bente, G. 136, 151

- Bentler, P. M. 185, 199, 201
 Berdie, D. R. 278, 291, 307, 319
 Berg, I. A. 229, 292
 Bergan, J. R. 35, 61
 Bergermeier 104, 125
 Bergler, R. 154, 192, 201
 Berlyne, D. E. 201
 Berk, R. A. 288, 292
 Berman, J. S. 105f, 125
 Bernard, J. L. 204
 Bettinghaus, E. P. 201
 Bien, W. 57, 61
 Binder, J. 286f, 292, 307, 319
 Bingham, W. 333, 356
 Binswanger, L. 394
 Birch, D. 201
 Birdsall, T. G. 125, 131
 Birdwhistell, R. L. 100, 125
 Bishop, F. M. 334, 357
 Bishop, Y. M. 29, 61
 Black, H. K. 201
 Black, J. L. 127, 130
 Blakeney, R. N. 329, 356
 Blalock, H. M. Jr. 43, 44, 61
 Block, J. 125, 136, 140, 141, 142, 143, 144, 145, 147, 148, 149, 201, 229, 292
 Blumenthal, A. L. 82, 125
 Bobbitt, R. G. 202
 Bock, R. D. 17, 61
 Böcker, W. 334, 356
 Böttgen, F. 16, 61
 Böttger, P. 397
 Bogo, N. 200
 Bohrstedt, G. W. 43, 61
 Bois, J. 71, 131
 Bokander, I. 202
 Bolstad, O. D. 121, 123, 125, 128
 Bolton, D. L. 338, 356
 Boltin, N. 207
 Borden, R. 295
 Borg, I. 6, 7, 61, 104, 125
 Borgatta, E. F. 102, 125
 Boring 91, 125
 Bortz, J. 192, 202, 206
 Boruch, R. F. 56, 58, 61
 Bousfield, W. A. 202
 Boyer, E. G. 13, 73
 Braaten, K. 297
 Bradburn, N. M. 229, 247, 265, 267, 270, 272f, 278, 283f, 284, 292, 301, 307, 320
 Bradley, P. 293
 Brandt, L. W. 187, 202
 Brannigan, C. R. 62
 Braun, P. 90, 125
 Bredenkamp, J. 16, 47, 62
 Brennan, R. L. 32, 33, 62
 Brewer, W. F. 202
 Brinton, J. E. 202
 Brod, Diane 178, 202
 Broden, M. 122, 125
 Brown, J. D. 212
 Brown, R. 191, 202
 Bross, I. 47, 62
 Broughton, A. 140, 153
 Brown, A. L. 97, 125
 Bruner, J. S. 104, 125
 Brunswik, E. 110, 125
 Buchner, D. 285, 287, 292
 Bühler, C. 365, 366, 381, 394
 Bühler, K. 348, 356
 Burger, G. K. 202
 Burgess, E. W. 375, 395
 Burisch, M. 225, 226, 241, 292
 Burland, S. 130
 Burnhill, P. 295
 Burns, E. 145, 149
 Burns, R. 173, 202
 Butler, J. M. 136, 142, 146, 147, 149
 Butler, L. 98, 130
 Butler, R. P. 284, 292
 Bynner, J. 173, 177, 202
 Cahalan, D. 249, 292
 Calder, P. H. 320
 Caldwell, B. M. 13, 62
 Callahan, D. M. 320
 Cameron, R. 130
 Campbell, D. T. 16, 51, 52, 53, 54, 55, 56, 57, 58, 60, 62, 74, 125, 133, 267, 292, 380, 389, 395
 Cannell, C. F. 227, 236, 238, 241, 243, 245f, 247, 252, 253, 256, 263, 265, 268, 270, 272, 276, 287, 292, 333, 357
 Canter, D. 7, 62
 Cantril, H. 254, 257, 266, 271, 299
 Caplan, S. W. 140, 149
 Carl, W. 229, 292, 293
 Carpenter, E. H. 307, 319
 Carroll, J. B. 163, 174, 202
 Carroll, R. M. 202
 Carroll, S. J. 247, 293
 Carter, D. J. 185, 191, 220
 Carter, R. F. 202
 Cartwright, D. S. 144, 149
 Cartwright, R. D. 140, 149
 Casey, M. 293
 Cash, W. B. 241, 300
 Cassel, R. N. 202
 Castellan, N. J. Jr. 111, 125
 Cataldo, E. F. 277, 292
 Cattell, R. B. 105, 125, 135, 145, 147, 149, 225, 292, 299, 325, 356, 392, 395

- Cavanaugh, J. C. 98f, 125
 Cavior, N. 119, 122, 125
 Centra, J. A. 53, 62
 Chaffee, S. H. 202
 Champneys, F. H. 365, 395
 Chapman, J. P. 105, 126
 Chen, Kathleen 178, 202
 Chapman, L. J. 105, 125
 Chase, P. H. 140, 149
 Chetwynd, S. J. 73
 Chi, P. L. 107, 126
 Chiba, Y. 215
 Child, I. L. 75
 Chipman, S. 126, 132
 Chmura, Kathy J. 212
 Christensen, A. 126, 128
 Chicchetti, D. V. 26, 62
 Ciminero, A. R. 117, 121, 126, 129
 Clanton, E. S. 296
 Claparede, E. 91, 126
 Clark, H. H. 203
 Clark, M. L. 203
 Clark, S. 184, 186, 220
 Clark, V. A. 186, 203
 Clarke, A. H. 2, 62
 Clauss, G. 249, 292
 Cleary, T. A. 48, 62
 Clevenger, T. 203
 Cliff, N. 187, 203, 229, 292
 Coan, R. W. 225, 293
 Cochran, W. G. 16, 48, 62
 Cohen, J. 30, 31, 33, 47, 62, 65, 143, 144, 149, 151
 Cohen, R. 105, 126, 229, 293
 Cohen, S. H. 48, 67
 Comte, A. 79, 82, 126
 Conger, A. J. 31, 56, 62
 Cook, S. W. 67, 180, 221
 Coombs, C. H. 107, 109, 126
 Corby, G. 299
 Corsini, R. J. 140, 149
 Cosper, R. 356
 Costner, H. L. 43, 44, 57, 63
 Cottrell, L. S. 125
 Cowen, L. 203
 Coyle, B. W. 71
 Coyle, F. A. 144, 149
 Coyne, L. 182, 203
 Craighead, W. E. 126, 128
 Crampton, C. W. 346, 356
 Cranach, M. v. 2, 4, 11, 12, 13, 63, 68
 Crapsi, L. A. 212
 Creelman, M. B. 203
 Crockett, W. H. 171, 203
 Cronbach, L.J. 41, 63, 65, 104, 126, 144, 145, 149, 188, 203, 224, 226, 229, 231, 289, 293
 Crutschfield, R. S. 293
 Cureton, E. E. 36, 63
 Dahme, B. 392, 395
 Dailey, C. A. 375, 395
 Damarin, F. 233, 293
 D'Andrade, R. G. 102f, 104, 108, 126
 Danziger, K. 91, 126
 Darnell, D. K. 172, 186, 190, 203
 Das Rhea, S. 167, 210
 Darwin., C. 365, 395
 Davis, A. 371, 395
 Davis F. B. 49, 63
 Dawes, R. M. 126, 188, 203
 De Boor, C. 370, 395
 De Burger, R. A. 203
 Deese, J. 184, 203
 Deming, W. E. 16, 63
 Denmark, F. L. 170, 203
 Deo, P. 146, 149
 Deters, H. 352, 356
 Deutsch, M. 67
 Deutschmann, P. J. 203
 De Waele, J. P. 375, 395
 Diab, L. N. 203
 Diament, C. 71, 126, 132
 Dice, L. R. 35, 63
 Dick, W. 181, 204
 Dicken, C. F. 203
 Dickman, H. R. 63
 Dickson, J. P. 283, 285, 293
 Dickson, W. J. 324, 359
 Diehl, B. 184, 188, 204
 Dietz, A. 127
 Dijkstra, W. 265, 297
 Dillehay, R. 258, 290, 293
 Dilthey, W. 362, 395
 Dilts, M. 204
 Dimock, H. S. 346, 357
 Dittrich, A. 199, 201
 DiVesta, F. 181, 204
 Dörner, D. 91, 126
 Dohrenwend, B. S. 299, 333, 359, 384, 399
 Dollard, J. 371, 375, 377, 395
 Donahoe, J. W. 203
 Dorsch, F. 135, 149
 Doyle, K. C. 290, 293
 Dudycha, A. L. 71
 Dukes, W. F. 364, 395
 Duncan, O. D. 43, 63
 Duncan, S. 100, 126
 Duncan, S. Jr. 2, 14, 64
 Duncker, K. 91, 226
 Dyke v. & Moore, B. V. 356
 Dymond, R. F. 140, 144, 145, 146, 147, 149, 152
 Ebel, R. L. 37, 38, 64
 Eber, W. H. 292
 Echelmeyer, L. 325, 357
 Edgell, S. L. 240, 289, 291
 Edinger, L. J. 370, 395

- Edwards, A. L. 224, 228f, 251, 293
 Egan, J. P. 107, 109, 127
 Ehlers, T. 230, 242, 261, 293
 Eisenman, R. 204
 Eiser, J. R. 180, 204
 Ekman, G. 168f, 211, 336, 357
 Ekehammar, B. 148, 150
 El-Batrawi, S. 144, 152
 Ellgring, J. H. 2, 62
 Elliott, L. L. 199, 204
 Ellis, A. 245, 293
 Endicott, J. 65
 Endler, N. S. 204
 Ends, E. J. 147, 150
 Engel, M. 150
 Engquist, G. 11, 71, 112, 127, 131
 Erbslöh, E. 310, 319, 322, 357
 Epperson, W. V. 284, 293
 Erdos, P. L. 242, 247, 262, 279f, 281, 283, 287, 293
 Erickson, M. 201
 Ericsson, K. A. 91, 114, 116, 127
 Erikson, H. E. 370, 395
 Ertel, S. 164, 167, 171, 172, 173, 185, 189, 204
 Ervin-Tripp, S. M. 205
 Esser, H. 312, 319
 Estep, M. F. 324, 356
 Etkind, A. M. 205
 Everitt, B. S. 32, 64, 65
 Everett, A. V. 168, 205
 Eysenck, H. J. 226f, 227, 246, 259, 293, 353, 357
 Fagan, J. 147, 150
 Falkenberg, S. D. 186, 201
 Falthzik, A. M. 247, 293
 Farr, S. D. 173, 210
 Faßnacht, G. 2, 11, 64
 Fear, R. A. 334, 357
 Feenstra, H. J. 206
 Feger, B. 88, 127
 Feger, H. 1, 2, 10, 11, 13, 18, 42, 43, 53, 57, 64, 71, 88, 113, 127, 132, 193, 205, 224, 294, 388, 395
 Feldt, L. S. 290, 301
 Fertig, J. 184, 219
 Fiedler, F. E. 140, 150
 Field, J. 202
 Fienberg, S. 61
 Filipp, S. H. 148, 150
 Finke, H. O. 295
 Finley, J. R. 205
 Finn, R. H. 40, 64
 Finke, R. A. 95, 127
 Fischer, G. 227, 231, 266, 287, 294
 Fischer, G. W. 48, 64
 Fischer, G. H. 148, 150
 Fischer, K. 354, 360
 Fishbein, M. 10, 64, 170, 171, 172, 176, 217, 220
 Fiske, D. W. 14, 51, 52, 53, 54, 55, 56, 57, 58, 60, 63, 64, 229, 267, 292, 294, 301
 Fisseni, H. J. 289, 294, 319, 347, 357
 Fixen, D. L. 119, 127
 Flade, A. 192, 205
 Flavell, J. H. 98, 127, 205
 Fleiss, J. L. 32, 33, 34, 35, 38, 46, 47, 64, 341, 357
 Fode, K. 206
 Fodor, J. A. 205
 Förster, P. 289, 294
 Fang, C. 36, 73
 Ford, L. H. 180, 205, 212
 Forthman, J. H. 198, 205
 Franke, J. 192, 206
 French, G. M. 185, 211
 Friedman, C. J. 206
 Foster, D. F. 399
 Foster, S. L. 68, 119f, 123f, 127, 129
 Fowler, R. D. 144, 149
 Fox, W. C. 127, 131
 Frank, G. H. 144, 146, 150
 Frawley, W. H. 36, 72, 72
 Freeman, W. 71
 Frenz, H. G. 2, 4, 11, 13, 65
 Freud, A. 370
 Freud, S. 368, 371, 378, 395
 Frey, D. 310, 319
 Frey, S. 11
 Frick, R. 26, 65
 Friedman, I. 140, 147, 150
 Friedrich, W. 241, 244f, 286, 289, 294
 Friedrichs, H. 325, 357
 Friedrichs, J. 5, 65, 227, 241, 245ff, 252, 254, 256, 263, 271, 294, 302, 319, 379, 395
 Frisbie, B. 248, 294
 Froehlich, C. P. 333, 357
 Froehlich, R. E. 357
 Frohburg, I. 136, 137, 138, 139, 144, 146, 147, 150
 Fuchs, A. 171, 183, 187, 188, 193, 194, 206
 Fürntratt, E. 229, 294
 Fuller, C. 283, 294, 307, 319
 Futrell, C. M. 284, 294
 Gärling, T. 168, 206
 Galanter, E. 14, 70
 Galtung, J. 255, 294

- Gardiner, D. H. 211, 221
 Gardner, R. C. 199, 206, 218
 Garfield, S. L. 140, 150
 Garrison, K. R. 372, 396
 Gebhard, P. H. 357
 Gehlen, H. 228, 297
 Gerbasi, K. C. 229, 231, 300
 Gesemann, G. 363, 396
 Getzels, J. W. 233, 294
 Gibson, J. J. 93, 127
 Giesen, H. 372, 394
 Giles, H. 211
 Gilpin, A. R. 186, 206
 Ginsberg, G. P. 297
 Giorgi, A. 99, 127
 Girard, R. 293
 Glad, B. 370, 396
 Gladden, J. W. 206
 Glasgow, W. D. 90, 127, 131
 Gleser, G. C. 41, 42, 63, 61, 145, 149, 183, 188, 203, 293
 Godfrey, R. R. 166, 206
 Goldberg, R. 295
 Goldberger, A. S. 43, 66
 Golden, G. H. 211
 Goldfried, M. R. 121, 127, 185, 197, 206, 207
 Goldiamond, I. 127
 Golding, S. L. 56, 58, 65
 Goldman, R. D. 212
 Goldstein, H. 396
 Goode, W. J. 269f, 273, 275, 282, 285f, 294
 Gooding, C. T. 144, 253
 Goodman, L. A. 29, 36, 65
 Gordon, D. A. 293
 Gottfried 117, 127
 Gottmann, J. M. 392, 396
 Gorospe, F. H. 206
 Graham, J. R. 140, 150
 Graumann, C. F. 1, 2, 3, 10, 11, 18, 65, 78, 82f, 127
 Gray, A. 281, 294
 Green, D. M. 28, 65
 Green, P. E. 168, 185, 197, 207
 Green, R. F. 207
 Greeno, J. G. 101, 128
 Griesemer, H. A. 288, 292
 Grigg, A. E. 207
 Grimm, G. 178, 191, 207
 Groeben, N. 317, 320
 Grooms, R. R. 215
 Gross, H. 324, 357
 Grubbs, F. E. 39, 65
 Grümer, K. W. 2, 5, 11, 13, 65, 128
 Gruhle, H. W. 396
 Gruson, L. 130
 Guertin, W. H. 142, 150
 Guilford, J. P. 66, 101, 106, 128, 145, 147, 150, 226, 249, 294
 Gullahorn, J. E. 280, 294
 Gullahorn, J. T. 280, 294
 Gulliksen, H. 109, 128, 182, 198, 207, 227, 294
 Gurwitsch, A. 92, 128
 Guski, R. 288, 295
 Gustad, J. W. 333, 356
 Guthrie, G. M. 140, 147, 150, 152
 Guttman, L. 6, 7, 8, 9, 66, 69, 266, 301
 Guttman, R. 6, 7, 8, 66
 Haan, N. 150
 Haase, H. 279, 282, 295
 Häcker, H. 230, 295
 Hagen, R. L. 121, 128
 Haggard, E. A. 37, 38, 66
 Haigh, G. V. 136, 142, 146, 149
 Haith, M. M. 365, 396
 Hall, C. 371
 Hall, R. V. 125i
 Hambitzer, M. 373, 396
 Harnikon, D. L. 178, 186, 207
 Hamilton, H. W. 207
 Hampel, R. 230, 295
 Hanlon, R. E. 147, 150
 Hannan, M. T. 43, 66
 Hannon, J. E. 204
 Hansen, M. H. 16, 66
 Harman, H. H. 167, 207
 Hart, J. T. 98, 228
 Hartemink, B. G. 359
 Hartley, J. 279, 281, 295
 Hartley, M. W. 140, 150
 Hartmann, H. 251, 295, 370
 Hase, H. 295
 Hasemann, K. 2, 66
 Hastorf, A. H. 207
 Hathaway, S. R. 242, 276, 295
 Hatt, P. K. 269f, 273, 275, 282, 285f, 294
 Hauser, R. M. 43, 66
 Haycock, V. 199, 220
 Hayes, D. P. 11, 66, 277f, 295
 Haygood, R. C. 184, 207, 218
 Heaps, R. A. 207
 Heard, W. G. 218
 Heberlein, T. A. 57, 60, 307, 319
 Heeler, R. M. 56, 71
 Heise, D. R. 43, 56, 66, 154, 167, 175, 188, 195, 199, 207
 Heller, D. 250, 295
 Hendel, D. D. 24, 66
 Henderson, E. H. 211
 Hendrick, C. 287, 295
 Hendrix, L. 290, 295
 Hennes, J. D. 220

- Hennig, W. 241, 271f,
 294, 295
 Hensley, W. E. 319
 Herbert, E. W. 122, 128
 Herder, J. G. 362, 396
 Herrmann, T. 178, 207,
 250, 295
 Hersen, M. 148, 151,
 364, 388, 389, 394, 396
 Heskin, K. J. 173, 207
 Heyns, R. W. 2, 66
 Hickey, M. E. 338, 356
 Higa, M. 184, 210
 Hildebrand, D. K. 36,
 66
 Hilden, A. H. 151
 Hill, R. J. 273, 301
 Hirschberg & Jones,
 L.E. 208
 Hoar, J. R. 207
 Hörmann, H. 154, 207
 Hoeth, F. 230, 242, 295
 Hoffmann, K. 208
 Hofstätter, P. R. 147,
 150, 190, 192, 208
 Hogg, J. 208
 Holeway, R. E. 197, 219
 Holland, P. 61
 Hollenbeck, A. R. 26,
 27, 28, 32, 45, 66
 Holley, J. W. 33, 66
 Holm, K. 226, 234, 241,
 243f, 249, 251, 264,
 267, 270, 272, 276, 295,
 296, 298
 Holtzmann, W. H. 396
 Holzkamp, K. 128
 Holzman, P. S. 182, 203
 Holz, W. 124
 Homzie, M. J. 208
 Hopper, D. 208
 Horman, M. 208
 Horn, J. L. 208
 Horner, A. J. 396
 Hornick, C. W. 225,
 296
 Hornke, L. 290, 296
 Hornstein, H. A. 320
 Horowitz, M. 71
 Hovland, C. I. 180, 208
 Howe, E. S. 179, 187,
 208
 Huber, H. P. 48, 66,
 392, 396
 Hubert, L.J. 32, 33, 36,
 44, 45, 54, 67
 Huck, S. W. 67
 Hübner, P. 298
 Humphries, D. A. 62
 Hundal, B. S. 146, 249
 Hunter, J. E. 48, 67
 Hunt, R. G. 212
 Hurley, J. R. 140, 152
 Hutwitz, D. 199, 208
 Hutwitz, W. N. 67
 Husek, T. R. 199, 200,
 208
 Husen, T. 339, 357
 Hutt, C. 2, 18, 67
 Hutt, S. J. 2, 18, 67
 Inbau, F. E. 332, 333,
 357
 Ingleby, J. D. 108f, 128
 Irish, D. P. 297
 Irle, M. 181, 208
 Irwin, L. 71
 Jackson, D. N. 55, 56,
 58, 67, 105, 129, 180,
 219
 Jacobs, P. I. 128, 133
 Jacobson, A. L. 43, 44,
 67
 Jaehner, D. 377, 396
 Jahoda, M. 2, 67
 Jakobowitz, L. A. 199,
 209, 211
 James, C. 212
 James, L. R. 296
 James, V. A. 324, 360
 James, W. 79, 81, 128
 Janke, W. 230, 287, 296,
 312, 319
 Jannsen, J. P. 230, 296
 Janson, S. 35, 67
 Jaspers, D. 375, 377,
 387, 396
 Jeanneret, P. R. 298
 Jenkins, E. 145, 149
 Jenkins, J. J. 209
 Jenkins, W. S. 329, 361
 Jernigan, L. R. 258, 290,
 293
 Jetzschmann, H. 241,
 296
 Joe, G. W. 67
 Jöreskog, K. G. 42, 43,
 56, 67, 68, 71, 74, 75,
 133, 304, 319
 John, D. 151
 Johnsgard, K. W. 217
 Johnson, A. 179, 205
 Johnson, C. A. 206
 Johnson, D. M. 106,
 120ff, 128
 Johnson, J. H. 308, 319
 Johnson, R. L. 209
 Johnson, R. M. 274, 292
 Johnson, S. M. 106,
 120ff, 128
 Johnson, W. R. 274, 296
 Jones, A. 143, 151, 296
 Jones, E. 369, 396
 Jones, J. M. 209
 Jones, R. R. 118, 129
 Jonsson, C. O. 241, 296
 Jordan, J. E. 15, 68
 Jung, C. G. 329, 357,
 370
 Kahn, R. L. 241, 245,
 253, 263, 272, 292, 296,
 333, 357
 Kahneman, D. 172, 174,
 178, 209
 Kalbermatten, U. 12, 68
 Kalinowsky-Czech,
 M. 229, 296
 Kalis, B. L. 151
 Kallabis, H. 241

- Kalleberg, A. L. 56, 68
 Kamp, K. 379, 395
 Kane, A. 146, 151
 Kane, R. B. 196, 209, 277, 296
 Kanfer, F. H. 90, 122, 129
 Kanowitz, J. 131
 Kant, I. 77, 79, 82, 129
 Kanungo, R. N. 209, 212
 Kaplan, K. J. 186, 198, 209
 Kaplitza, G. 241, 255, 260, 267, 278, 280, 296
 Kardiner, A. 371, 377, 396
 Karmasin, F. 241, 244f, 248, 252, 254, 257, 259f, 261f, 264, 266f, 270f, 273f, 278, 281, 287, 296
 Karmasin, H. 241, 244f, 248, 252, 254, 257, 259f, 266f, 270f, 273f, 278, 281, 287, 296
 Kashiwagi, S. 209
 Katkin, E. S. 105, 133
 Katz, D. 92, 129
 Katz, M. 209
 Kaufman, H. J. 209
 Kaufman, I. C. 13, 68
 Kavanagh, M. J. 60, 68
 Kaye, K. 35, 46, 68
 Kazdin, A. E. 122, 129, 364, 396
 Keehn, J. D. 215
 Keeser, W. 392, 396, 399
 Keil, A. 209
 Keil, C. 209
 Keil, W. 151, 230, 261, 296
 Kellstedt, L. A. 292
 Kelly, G. A. 142, 151
 Kelly, J. A. 210
 Kemmler, L. 325, 357
 Kemnitzer, L. S. 140, 151
 Kendall, M. G. 36, 68
 Kennedy, D. R. 141, 152
 Kenny, D. A. 68, 105f, 125
 Kent, R. N. 2, 68, 71, 117, 119f, 123f, 127, 129, 131f
 Kentler, H. 210
 Kepes, S. Y. 284, 296
 Kephart, N. 324, 357
 Kerkhoff, T. 17, 71
 Kerlinger, F. N. 151
 Kernoff, P. 202
 Kerrick, J. S. 203
 Kessen, W. 365, 396
 Keys, A. 46, 68
 Kihlberg, J. K. 46, 68
 Kilty, K. M. 210
 King-Fun Li, A. 210
 Kinsey, A. C. 257, 296, 323, 331, 332, 333, 334, 335, 343, 344, 346, 357, 384, 396
 Kipnowski, A. 397
 Kirby, D. M. 206
 Kirchner, F. Th. 380, 397
 Kiresuk, T. J. 392, 397
 Kirschhofer-Bozenhardt, A. v. 241, 258, 260, 267, 278, 280, 296
 Kish, G. B. 286f, 296
 Kish, L. 16, 68
 Kissel, E. 397
 Kitay, D. H. 369, 397
 Kjeldergaard, P. M. 184, 210
 Klages, L. 340, 357
 Klapper, J. T. 210
 Klauer, K. J. 49, 68, 141, 148, 151
 Klebelsberg, D. v. 328, 357
 Klein, D. 299, 333, 359, 384, 399
 Klemmack, D. L. 173, 177, 210
 Klinkhammer, F. 230, 291
 Kluckhohn, F. R. 5, 68
 Kluegel, J. R. 56, 68
 Knudsen, D. D. 283, 297
 Koch, G. G. 26, 35, 37, 47, 68, 69
 Köbler, V. 242, 295
 Köhler, W. 93, 129
 Koen, F. 210
 König, R. 2, 68, 241, 297
 Kogan, N. 133
 Kohli, M. 362, 372, 397
 Komorita, S. S. 210
 Koolwijk, J. v. 235, 240f, 274, 297, 303, 319
 Koomen, W. 265, 297
 Korman, M. 210
 Korner, A. F. 37, 42, 69
 Kornreich, B. 146, 151
 Kosslyn, S. M. 95, 127
 Kostić, D. 167, 210
 Kraemer, H. C. 37, 42, 69
 Kratochwill, T. R. 364, 388, 397
 Krause, I. 210
 Krause, M. S. 54, 69
 Krauss, H. 50, 74
 Kraut, A. I. 273f, 278, 297
 Kravetz, S. 219
 Kreipe, K. 324, 358
 Kreutz, H. 241f, 246, 251, 255, 257ff, 262ff, 267, 269f, 274f, 286, 297, 313, 319
 Krieger, M. H. 172, 180, 210
 Krippendorff, K. 17, 35, 38, 40, 69
 Kris, E. 370
 Kriz, J. 30, 69
 Kröber, W. 324, 358
 Kroes, W. H. 184, 210
 Krüger, P. 250, 295

- Kruglanski, A. W. 16, 69
 Kruskal, W. H. 29, 36, 65, 69
 Kubinieć, C. M. 173, 210
 Kuiper, P. C. 363, 370, 397
 Kumata, H. 168, 210
 Kuncel, R. B. 229, 297
 Kunkel, E. 334, 335, 336, 348, 358
 Kuusinen, J. 164f, 176, 199, 211
- Lachman, J. L. 99, 129
 Lachman, R. 99, 129
 Laing, J. D. 66
 Lalu, N. M. 43, 44, 67
 Lambert, W. E. 200
 Lambert, W. W. 75, 100, 129
 Lampl de Groot 370, 397
 Lana, R. E. 211
 Landis, D. 219
 Landis, J. R. 26, 35, 37, 65, 69
 Lane, S. T. M. 211
 Langer, I. 135, 151
 Lansing, J. B. 227, 236, 297
 Lantermann, E. D. 297
 Larkin, J. D. 62
 Laurent, A. 292
 LaVoie, A. L. 199, 201
 Lawlis, G. F. 40, 69
 Lawson, E. D. 211
 Laxer, R. M. 141, 152
 Lay, C. H. 105, 129
 Lazarsfeld, P. F. 247, 252, 297
 Lazier, G. A. 203
 Leach, G. M. 69
 Ledvinka, J. 358
 Lehr, U. 334, 342, 348, 358, 372, 397
- Leitenberg, H. 364, 388, 389, 397
 Lennertz, E. 288, 297
 Levin, J. 171, 211
 Levy, L. H. 146, 147, 151, 177, 210
 Levy, N. 185, 216
 Levy, P. 9, 69, 211
 Levy, S. 42, 69
 Lewin, K. 329, 358
 Lewis, O. 371, 397
 Libby, W. L. 184, 210
 Lichtenstein, E. H. 202
 Lieberman, D. A. 90, 129
 Lienert, G. A. 26, 33, 66, 69, 227, 266, 287, 297, 336, 358
 Light, C. S. 178, 211
 Light, R. J. 30, 31, 32, 33, 62, 69
 Lilli, W. 181, 211
 Lindsey, D. 295
 Linehan, M. M. 121, 127
 Linn, R. L. 48, 56, 62, 71, 74, 75, 133
 Linschoten, J. 229
 Linsky, A. S. 279, 283, 285, 287, 297, 307, 319
 Linton, R. 375, 397
 Lipinski, D. P. 119, 126, 129f
 Lipmann, O. 356
 Lisch, R. 30, 69
 Litt, E. N. 171, 211
 Litwak, E. 257, 297
 Livson, N. H. 144, 151
 Lobitz, G. K. 120, 128
 Loch, W. 370, 297
 Loewenstein, R. M. 370, 398
 Lohr, J. M. 166, 211
 Long, B. 178, 179, 211
 Longabaugh, R. 2, 14, 17, 20, 23, 70, 99, 130
 Lopez, F. M. 338, 354, 358
 Lord, F. M. 43, 70
- Lorr, M. 105, 130
 Louiselle, R. M. 221
 Lowy, D. G. 221
 Lu, E. 40, 69
 Lu, K. H. 38, 70
 Luce, R. D. 14, 70
 Lück, H. E. 207
 Lüdtke, H. 5, 65
 Lüer, G. 91, 111, 130
 Luria, Z. 214
 Lysterly, S. B. 36, 70
 Lyle, J. 211
- Maccoby, E. E. 241, 244, 248, 251, 259, 271f, 298
 Maccoby, N. 241, 244, 248, 251, 298
 Mac Kinney, A. C. 62, 68
 Maclay, H. 211
 Madden, J. E. 211
 Madow, W. G. 66
 Magnussen, D. 227, 298
 Magnusson, D. 168, 211, 337, 358
 Magnusson, F. 168, 220
 Maguire, T. O. 211
 Makohoniuk, G. 118, 130
 Maletzky, B. M. 121, 130
 Malmstrom, E. J. 185, 211
 Maltz, H. E. 178, 212
 Manis, M. 212
 Mann, F. 276, 286, 298
 Mann, J. H. 125
 Mann, R. D. 102f, 130
 Manz, W. 2, 13, 70
 Marabotto, C. 119, 122, 126
 Marcill J. C. 147, 148, 151
 Margis, P. 323, 358
 Markel, N. N. 184, 212

- Marks, E. S. 243, 272, 298
 Marks, I. 178, 212
 Marks, P. A. 140, 144, 149, 151
 Marquis, K. H. 292
 Martin, C. E. 296, 333, 357, 396
 Martinez, J. L. jr. 212
 Martinez, S. R. 212
 Maruyama, K. 183, 212
 Mash, E. J. 118, 130
 Mash, L. J. 130
 Maslow, A. 367, 398
 Mason, W. M. 267, 270, 272f, 278, 292
 Massarik, F. 367, 381, 394
 Matteson, M. T. 307, 320
 Mauldin, W. P. 243, 272, 298
 Mauxois, A. 377
 Maxwell, A. E. 37, 70
 May, W. H. 154, 165, 176, 215, 219
 Mayer, L. S. 43, 33, 70
 Mayerberg, C. K. 212
 Mayfield, E. C. 342, 353, 358
 Mayntz, R. 241, 251, 262, 298
 Mazlish, B. 370, 398
 McCall, G. J. 5, 6, 70
 McCallon, E. L. 212
 McClave, J. T. 74
 McCormick, E. J. 289, 298
 McDermott, P. A. 31, 70, 74
 McDonagh, E. C. 286, 298
 McElwee, J. D. 118, 130
 McFarlane, J. W. 323, 358
 McGinnies, E. 216
 McGrew, W. C. 2, 70
 McGuire, W. J. 380, 386, 398
 McKelvie, S. J. 198, 212, 298
 McKinley, J. C. 242, 276, 295
 McKinley, S. M. 29, 70
 McNair, D. M. 130
 McNaughton, J. F. 329, 336
 McNicol, D. 28, 70, 107, 130
 Mecham, R. C. 298
 Medley, D. M. 2, 13, 14, 70
 Meehl, P. E. 63, 226, 293
 Meek, E. E. 207
 Mees, U. 2, 71
 Meichenbaum, D. 97f, 130
 Meier, R. D. 334, 358
 Meisels, M. 180, 205
 Mellenbergh, G. J. 109, 129, 133, 359
 Meltzer, L. 66
 Messer, S. 212
 Messick, S. J. 187, 188, 212
 Metge, A. 82f, 130
 Metzger, W. 92, 130, 334, 358
 Metzler, J. 95f, 132
 Metzner, H. 276, 286, 298
 Meurer, K. 144, 145, 152
 Michel, L. 358
 Micko, H. C. 192, 212
 Middleton, D. 107, 133
 Mierke, K. 324, 358, 384, 398
 Miettinen, O. S. 398
 Mikula, G. 178, 191, 192, 212
 Milbrath, L. W. 292
 Miles, M. B. 320
 Miller, F. D. 320
 Miller, G. A. 212
 Miller, P. McC. 212
 Miller, R. L. 310, 316, 320
 Miller, S. 212
 Mills, D. H. 192, 213
 Mindak, W. A. 197, 213
 Minsell, W.-R. 136, 151
 Miron, M. S. 154, 161f, 182, 190, 213, 215
 Misch, E. 363, 398
 Mitchell, D. B. 95, 130
 Mitsos, S. B. 178, 191, 192, 213
 Mittenecker, E. 224f, 231, 233, 242, 298
 Mitts, B. 125
 Mitzel, H. E. 2, 13, 14, 70
 Moffatt, G. W. 353, 359
 Mogar, R. E. 178, 213
 Moore, B. V. 333, 356
 Moos, R. H. 121, 130
 Mordkoff, A. M. 185, 213
 Morimato, H. 213
 Morland, J. K. 220
 Morris, C. 215
 Moss, C. S. 213
 Mote, V. L. 71
 Mower White, C. J. 180, 204
 Mowrer, H. H. 135, 151
 Mucchielli, R. 241, 298
 Mueller, W. S. 213
 Münch, W. 241, 298
 Mulaik, A. 105, 130
 Munoz, S. R. 200
 Munroe, R. L. 20, 71
 Murray, E. J. 295
 Murray, H. A. 368, 377
 Mutherr, B. 75, 172, 213
 Nahinsky, I. D. 140, 151
 Nanda, H. 63, 293
 Narayana, C. L. 249, 298

- Natalicio, L. F. S. 166, 206
 Natsoulas, T. 90, 130
 Nay, W. R. 17, 71
 Naylor, J. C. 17, 71
 Nee, J. C. M. 65
 Neff, W. S. 143, 151
 Nelson, R. O. 119, 126, 129, 130
 Neufeld, R. W. 388, 398
 Neuringer, C. 178, 213
 Newcomb, T. M. 104, 131
 Newton, D. 11, 71, 112, 131
 Nicewander, W. A. 47, 71
 Nichols, H. J. 217
 Nichols, T. F. 144, 151
 Nickels, S. A. 179, 191, 213
 Nidorf, L.J. 171, 203
 Niederland, W. G. 369, 398
 Nisbett, R. E. 79, 131, 316, 320
 Nishimura, H. 216
 Noelle, E. 241f, 248, 250, 259f, 266f, 278, 298
 Noelle-Neumann, E. 239, 252ff, 262, 266, 268f, 271ff, 275, 278, 298
 Nordenstreng, K. 164f, 168, 172, 175, 213, 214
 Norman, W. T. 105, 131, 181, 214
 Notarius, C. 392, 396
 Novick, M. R. 43, 70
 Nowakowska, M. 229, 233, 298
 Nunnally, J. C. 146, 152
 Nußbaum, A. 42, 71
 Nuttin, J. 92, 131
 Obst, G. H. 372, 398
 O'Connell, E. J. 62
 O'Connor, J. 147, 150
 O'Donovan, D. 191, 214
 Oetting, E. R. 191, 214
 Olbrich, M. 349, 359, 398
 Oldfield, R. C. 324, 359
 O'Leary, K. D. 17, 124, 129, 131, 132
 Oles, H. J. 182, 214
 Olmedo, E. L. 212
 Olsson, U. 213
 O'Mally, J. M. 217
 Ono, H. 207
 Oppenheim, A. N. 241, 264, 298
 Orlik, P. 167, 178, 191, 214
 Orne, M. T. 333, 359
 Osgood, C. E. 154f, 172ff, 188ff, 200, 214, 215, 218, 219, 250, 261, 298
 Osipow, S. 215
 Osterland, M. 372, 398
 Otis, J. L. 201
 Overall, J. E. 39, 71
 Oyama, T. 214, 218
 Padden, D. 208
 Page, C. W. 147, 150
 Paivio, A. 215
 Paponia, N. 71
 Papousek, H. 2, 73
 Parkinson, C. N. 324, 359
 Parloff, M. B. 147, 148, 153
 Parouson, B. S. 392, 398
 Paskewitz, D. A. 333, 359
 Passini, F. T. 105, 131
 Pastore, R. E. 107, 131
 Patterson, G. R.
 Paul, G. L. 118, 128, 131
 Paul, S. 372, 398
 Pauleinkhoff, B. 398
 Pauli, R. 356
 Pauling, F. J. 211
 Payne, S. L. 247, 249, 251ff, 258, 261, 263f, 266, 282, 298
 Peabody, D. 178, 215
 Peak, H. 2, 71
 Pearson, K. 34, 71
 Peck, R. C. 284, 293
 Peckham, S. 201
 Perkins, H. V. 140, 152
 Perlmutter, M. 98f, 126
 Perreault, W. D. 269, 277, 279, 282, 290, 299
 Petermann, F. 364, 380, 381, 387, 388, 392, 397, 398, 399
 Petermann, U. 381, 399
 Peters, D. J. 73
 Peters, L. H. 260, 301
 Peterson, A. O. D. 152
 Peterson, D. R. 131
 Peterson, W. W. 104, 131
 Pettersson, T. 213
 Pfahler, G. 323, 359
 Phillips, B. S. 241, 254, 257, 259f, 268, 270, 272, 299
 Phillips, E. L. 127, 152
 Piaggio, L. 181, 215
 Pickett, L. 202
 Pilkington, C. W. 131
 Pilliner, A. E. G. 37, 70
 Plutchik, R. 215
 Podgorny, P. 131
 Polansky, N. 5, 71
 Pomeroy, J. E. 296
 Pomeroy, W. B. 333, 357, 396
 Pongratz, L. 323, 359
 Pool, J. 11, 65
 Pope, B. 208
 Pope, H. 297
 Postman, L. 110, 131
 Powell, G. E. 73
 Prager, R. A. 140, 150
 Presly, A. S. 173, 177, 215

- Presser, S. 312, 320
 Preyer, W. 365, 399
 Price, J. M. 47, 71
 Price, R. H. 107, 131
 Priest, P. N. 178, 215
 Procter, C. H. 47, 61
 Prothro, E. T. 215

 Quarter, J. 141, 152

 Raab, E. 263, 266, 299
 Rachlin, H. 122, 131
 Radford, J. 90, 131
 Raiford, A. 144, 152
 Rajaratnam, N. 38, 63, 72, 206, 293
 Rao, P. V. 74
 Rao, V. R. 207
 Rapaport, D. 71
 Rappaport, J. 204
 Rasmussen, V. 399
 Ray, M. L. 56, 72
 Ray, W. S. 152
 Rechetnick, J. 359
 Redfield, J. 118, 131
 Reece, M. W. 215
 Reed, T. R. 215
 Reid, J. B. 18, 72, 118f, 132
 Reid, J. E. 332, 333, 357
 Renzaglia, G. A. 120, 132
 Revenstorff, D. 167, 171, 174, 180, 187, 188, 199, 216, 392, 399
 Revie, V. A. 140, 152
 Richardson, J. T. E. 94, 132
 Richardson, S. A. 241, 299, 333, 359, 384, 399
 Richman, C. L. 95, 130
 Richter, H. J. 241, 243, 247, 249, 266, 269, 273, 276, 279ff, 284f, 286ff, 299
 Riley, R. T. 203

 Rindner, R. J. 112, 131
 Ring, E. 256, 281, 299
 Risley, T. R. 388, 394, 399
 Roberts, R. R. Jr. 120, 132
 Robinson, P. W. 399
 Robinson, R. 219
 Rock, D. A. 48, 72
 Roessler, E. B. 36, 60
 Roethlisberger, F. J. 324, 359
 Rogers, A. H. 140, 152
 Rogers, C. R. 140, 146, 147, 152
 Rogers, T. B. 229, 299
 Rogot, E. 47, 72
 Rohrachner, H. 82, 132
 Rohrmann, B. 229, 249, 264, 295, 299
 Roll, S. 216
 Romanczyk, R. G. 17, 18, 72, 118, 132
 Romein, J. 375, 377, 399
 Romney, D. 173, 177, 202
 Rorer, L. G. 231, 299
 Roscoe, A. 278, 300
 Rosenbaum, L. L. 169, 216
 Rosenbaum, W. B. 216
 Rosenblum, A. L. 286, 298
 Rosenblum, L. A. 13, 68
 Rosenmeier, H. P. 334, 359
 Rosenthal, H. 66
 Rosenthal, R. 16, 72, 120, 132
 Roslow, S. 244, 247, 257, 260, 266, 299
 Rosnow, R. L. 16, 71, 120, 132
 Ross, B. M. 216
 Ross, J. 185, 216
 Ross, M. 317, 320
 Rothenberg, A. 297
 Rothman, A. I. 216

 Rowe, P. M. 329, 359
 Rubin, M. 140, 152
 Robinson, R. 66
 Rudinger, G. 13, 42, 72, 104, 132
 Rugg, D. 254f, 257f, 261, 266, 271, 299
 Rugg, H. 104, 132
 Ruggels, W. L. 202
 Ruppel, M. 364, 392, 399
 Ruppenthal, G. C. 72
 Russell, W. A. 209
 Rydell, S. T. 178, 216
 Rytten, B. 212

 Saari, B. B. 71
 Sackett, G. P. 18, 19, 72
 Sagara, M. 216
 Salapatek, P. H. 365, 396
 Salber, W. 325, 359
 Sappenfield, B. R. 146, 152
 Sarason, I. G. 308, 319
 Schäfer, B. 171, 184, 188, 193, 194, 199, 204, 216
 Schapp, W. 92, 132
 Scharnweber 374, 399
 Scheele, B. 317, 320
 Schefflen, A. E. 100, 132
 Scheier, I. H. 225, 299
 Scheirer, C. J. 107, 131
 Schenck, E. A. 17, 71
 Scherer, K. R. 2, 72, 306, 320
 Scheuch, E. 322, 359
 Seheuch, E. K. 223, 228, 241, 245, 251, 261, 286, 299, 300, 302, 305, 320
 Schiavo, R. S. 320
 Schick, A. 191, 192, 216
 Schlosberg, H. 164, 216
 Schludermann, E. 216
 Schludermann, S. 178, 216

- Schmidt, H. D. 325,
334, 347, 359, 360, 367,
399
- Schmitt, N. 56, 58, 60,
72
- Schneider, J. 229, 233,
264, 300
- Schneider-Düker, M.
264, 300
- Schön, G.-H. 152
- Schoenberg, R. 57, 63
- Schönflug, W. 198, 216
- Schoggen, P. 20, 72,
100, 124
- Schonfeld, W. A. 346,
359
- Schraml, W. 325, 333,
334, 359, 370, 399
- Schramm, W. 168, 210
- Schreiber, K. 241, 300
- Schriesheim, C. 229, 300
- Schriesheim, J. 300
- Schucany, W. R. 36, 72
- Schümer, R. 105, 126,
132
- Schuh, A. J. 217
- Schuller, A. 334, 359
- Schulter, G. 178, 191,
192, 212
- Schulz, R. 296
- Schulz, W. 356
- Schulz v. Thun, F. 135,
151
- Schumann, H. 312, 320
- Schutz, W. C. 29, 72
- Schwartz, C. G. 5, 72
- Schwartz, M. S. 5, 72
- Schwartz, R. D. 74, 133
- Schwarzer, R. 308, 320
- Schwenkmezger, P. 295
- Schyberger, B. W. 245,
300
- Scott, C. 307, 320
- Scott, R. A. 60
- Scott, W. A. 16, 30, 73
- Scupin, E. 365, 399
- Scupin, G. 365, 399
- Sears, R. R. 368, 399
- Sechrest, L. 74, 133
- Seeman, W. 140, 151
- Seidman, E. 56, 58, 65
- Selg, H. 2, 71
- Seligman, E. 373, 399
- Selvage, R. 37, 73
- Semmel, M. 1. 26, 65
- Settle, R. B. 35, 74
- Seyfried, B. A. 295
- Shapiro, M. B. 148, 152
- Sharma, S. N. 280, 300
- Shaw, M. E. 179, 213
- Shaw, R. 378
- Sheatsley, P. B. 241, 300
- Shell, S. A. 217
- Shepard, R. N. 95f, 131,
132
- Shepherd, I. L. 140, 152
- Sherif, M. 188, 208
- Sherman, R. E. 392, 397
- Sherry, P. 140, 152
- Sheth, J. 278, 300
- Shikiar, R. 168, 172, 217
- Shinn, M. W. 365, 399
- Shirk, E. J. 203
- Shlien, J. M. 140, 152
- Shontz, F. C. 140, 152
- Shrout, P. E. 38, 65
- Sicoly, F. 317, 320
- Sieber, M. 287, 289, 292,
300, 307, 19
- Sieveking, N. A. 296
- Simkins, L. 119, 132
- Simon, A. 13, 73
- Simon, H. A. 91, 14,
116, 127
- Simons, G. 2, 73
- Simons, J. L. 5, 6, 70
- Simonton, D. K. 386,
399f
- Simpson, R. H. 229, 00
- Sines, J. O. 217
- Singer, R. D. 217
- Singh, Y. P. 280, 00
- Singleton, W. T. 73
- Sixtl, F. 231, 300
- Skinner, B. F. 89f, 93,
100, 132
- Slobin, D. I. 205
- Smith, E. R. 320
- Smith, F. V. 207
- Smith, G. 220
- Smith, R. G. 191, 217
- Snider, J. G. 154, 217
- Snyder, F. 171, 172,
175, 179, 217
- Snyder, F. W. 217
- Snyder, W. U. 152
- Sörbom, D. 43, 56, 67,
68, 304, 319
- Solarz, A. K. 217
- Solle, R. 210
- Somers, R. H. 73
- Sommer, R. 196, 217
- Sorembe, V. 41, 73
- Soudijn, K. A. 334, 359
- Spearman, C. 34, 36, 73
- Spiegel, B. 360
- Spinner, B. 98, 124
- Spitzer, R. L. 65
- Spoerer, E. 284, 400
- Spranger, E. 363, 400
- Spriegel, W. R. 324, 360
- Springbett, B. M. 217
- Staats, A. W. 166, 205,
218
- Staats, C. K. 218
- Stäcker, K. H. 250, 295
- Stahlberg, G. 213
- Stanley, J. C. 16, 57, 58,
62, 73, 107, 133, 380,
389, 395
- Starr, D.J. 105, 133
- Steinkamp, S. W. 329,
360
- Steller, M. 144, 145, 152
- Stephenson, W. 135,
140, 142, 153
- Stern, C. 365, 400
- Stern, W. 85, 133, 323,
324, 356, 360, 365, 400
- Steward, C. J. 241, 300
- Steward, R. B. 73
- Steward, T. R. 133
- Stewart, R. A. 36, 73

- Stollberger, R. 242, 246, 275, 300
 Stover, D. O. 204
 Strahan, R. 229, 231, 300
 Straka, J. 146, 151
 Stricker, G. 178, 218
 Stricker, L. J. 105, 133
 Stroebe, W. 180, 204
 Strong, E. R. 289, 300
 Stroschein, F. R. 241, 244ff, 250, 254, 256, 261, 266, 268f, 274, 285, 301
 Stroud, T. W. F. 48, 73
 Subotnik, L. 146, 153
 Suchman, E. A. 266, 301
 Suci, G. 168, 188, 215
 Suci, G. J. 154, 209, 218, 298
 Sudman, S. 229, 247f, 265, 283f, 292, 294, 301, 307, 320
 Süllwold, F. 228, 301
 Suk, J. M. 310, 320
 Summers, G. F. 73
 Susman, E. J. 26, 73
 Sutcliffe, J. P. 48, 73
 Swaminathan, H. 387, 400
 Swan, J. E. 284, 294
 Swets, J. A. 28, 65, 107, 133
 Tack, W. H. 400
 Taft, R. 329, 360
 Tagiuri, R. 104, 125
 Taietz, P. 283, 301
 Taine, H. M. 365, 400
 Tajfel, H. 181, 218
 Takahashi, S. 218
 Tamulonis, V. 292
 Tanaka, Y. 162, 170, 172, 215, 218
 Tannenbaum, P. 215
 Tannenbaum, P. H. 154, 199, 204, 218, 298
 Tansill, R. 212
 Taplin, P. S. 118f, 133
 Tatsuoka, M. M. 292
 Taubert, H. 296
 Taylor, C. L. 184, 218
 Taylor, D. M. 146, 153, 206, 218
 Taylor, H. F. 219
 Taylor, R. E. 204
 Taylor, W. L. 36, 73
 Terborg, J. R. 260, 301
 Terwilliger, R. F. 185, 202, 219
 Tesser, A. 50, 74
 Thackray, R. I. 333, 359
 Tholey, V. 229, 301
 Thomae, H. 2, 13, 74, 88, 133, 322, 323, 331, 333, 343, 348, 353, 360, 375, 397, 400
 Thomas, W. I. 371, 378, 400
 Thompson, C. 212
 Thoms, K. 322, 360
 Thorndike, E. L. 104, 133
 Thorne, F. C. 334, 360
 Thornton, G. C. 220
 Thumin, F. J. 201
 Tiedemann, D. 365, 400
 Timaeus, E. 207
 Tinsley, H. E. A. 26, 40, 74
 Titchener, E. B. 81, 133
 Titscher, S. 241f, 246, 251, 255, 257ff, 262ff, 267, 269f, 274f, 286, 297
 Tittle, C. R. 273, 301
 Tobacyk, J. J. 140, 153
 Tolman, E. C. 110, 131
 Tränkle, U. 223, 282, 285, 301
 Trankell, A. 324, 350, 360
 Triandis, H. C. 168, 192, 199, 219
 Triebe, J. K. 354, 355, 360
 Triebe, K. 320
 Trippi, R. R. 35, 74
 True, J. E. 284, 296
 Trumbo, D. 341, 353, 360
 Trush, R. S. 146, 153
 Tucker, L. R. 56, 58, 74
 Turner, C. 229, 301
 Turner, R. H. 140, 153
 Turvey, M. T. 184, 219
 Tversky, A. 126
 Tzeng, O. C. S. 165, 171, 172, 176, 183, 199, 219
 Ulich, E. 354, 355, 360
 Ulrich, L. 341, 353, 360
 Underwood, W. L. 220
 Undeutsch, U. 325, 326, 334, 360
 Utz, H. 295
 Van Atta, R. E. 140, 153
 Van der Kamp, L. J. T. 109, 129, 133
 Vanderlippe, R. H. 140, 153
 Van Meter, D. 107, 133
 Vaught, G. M. 140, 153
 Vegelius, J. 33, 67, 74
 Verinis, J. S. 216
 Verner, H. W. 292
 Vernon, Ph. E. 338, 360
 Vidal, J. J. 183, 197, 219
 Villamin, A. C. 206
 Vincent, R. A. 144, 153
 Vitale, J. 62, 74
 Vogel, B. 392, 399
 Voyce, C. D. 180, 219
 Wackerly, D. D. 31, 32, 74
 Wagner, R. 341, 361
 Wahl, D. 316, 320
 Walker, B. A. 219

- Walker, R. N. 153
 Wall, D. D. 179, 209
 Wallbott, H. G. 26, 32, 37, 61
 Walther, E. H. 352, 361
 Ware, E. E. 162, 170, 211, 215, 220
 Warr, P. B. 199, 220
 Warren, J. T. 66
 Washington, W. N. 178, 220
 Waskow, J. E. 147, 148, 153
 Watkins, M. W. 31, 70, 74
 Watson 89
 Wattawa, S. 290, 293
 Webb, E. J. 74, 120, 133
 Weick, K. E. 2, 20, 74
 Weigel, R. G. 220
 Weigel, V. M. 220
 Weimer, J. 208
 Weinreich, U. 220
 Weiss, D.J. 24, 26, 40, 74
 Weksel, W. 220
 Wellek, A. 326, 339, 352, 353, 355, 361
 Wells, F. L. 104, 133
 Wells, H. G. 378
 Wells, W. D. 220
 Welzel, U. 328, 350, 361
 Werner, J. 37, 38, 74
 Wertheimer, M. 73, 93, 133
 Werts, C. E. 42, 43, 48, 56, 71, 74, 75, 110, 133
 Whaley, F. 71
 Wheaton, B. 43, 75
 Whelan, P. 75
 White, G. 122, 128
 White, P. 90, 133
 White, R. W. 400
 Whiting, B. B. 100, 133
 Whiting, J. W. 100, 133
 Whiting, J. W. M. 17, 75
 Whitney, D. R. 290, 301
 Whyte, W. F. 5, 75, 332, 361
 Wichmann, U. 295
 Wickens, D. D. 184, 220
 Wieczorek, R. 205
 Wieken, K. 241, 276, 285, 286f, 301, 305, 320
 Wieken-Mayser, M. 297
 Wiendieck, G. 310, 319
 Wieser, I. 400
 Wiggins, N. 170, 171, 172, 175, 176, 179, 208, 217, 220
 Wilbur, P. H. 144, 153
 Wilcox, R. C. 220
 Wilder, D. 113, 133
 Wildman, R. C. 284, 301
 Wildman, R. W. 220
 Wiley, D. E. 43, 75
 Wiley, J. A. 43, 75
 Wiley, L. 329, 361
 Wilk, G. 241, 301
 Williams, J. E. 220
 Williams, W. S. 170, 220, 221
 Willick, D. H. 271, 278, 301
 Wilson, R. N. 368, 400
 Wilson, T. D. 79, 131, 316, 320
 Wilson, T. P. 36, 75
 Winer, B. J. 38, 75
 Winograd, E. 221
 Wittenborn, J. R. 135, 147, 148, 153
 Wittrock, M. C. 208
 Wohlfart, E. 361
 Wolf, G. 66
 Wolf, M. M. 127, 394, 399
 Wolfe, L. A. 334, 361
 Wolfenstein, E. v. 370, 400
 Wolfson, A. D. 297
 Wolins, L. 56, 62, 68
 Wonnacott, E. J. 206
 Woodward, J. A. 67
 Woog, P. C. 146, 153
 Wottawa, H. 226f, 262, 266, 268, 287, 301
 Wright, H. F. 2, 18, 20, 75, 377, 394
 Wright, O. R. 353, 355, 361
 Wright, P. 263f, 281, 301
 Wulfeck, H. 299
 Wundt, W. 80, 82f, 83, 134
 Wyckoff, D. 293
 Wylie, R. C. 138, 144, 145, 146, 153
 Wynd, W. 293
 Yamamoto, K. 216
 Yates, F. 16, 75
 Young, D. D. 221
 Younger, M. S. 43, 44, 70
 Zander, A. F. 2, 66
 Zaniecki, F. 371, 378, 400
 Zavalloni, M. 180, 221
 Zax, M. 178, 211, 218, 221
 Zehnpfennig, H. 300
 Ziller, R. C. 211
 Zippel, B. 221

Sach-Register

Anamnese 322

Antworttendenzen 230, 234, 260f

Bedeutung 164

affektive - 164

konnotative - 164

denotative - 164

Befragung 302ff, 321f

Begriffsbestimmung 301f

Fehlerquellen bei - 306, 310f

-, Methodenvarianten 303f

mündliche - 308f

Realkontakt - 313f

schriftliche - 305f

telefonische - 314f

- von Kindern 315f

Behaviorismus 89ff

Beobachtung(en) 1ff

-, Auswahl des zu beobachtenden
Verhaltens 18ff

-, Auswahl und Schulung von
Beobachtern 17f

Auswahl von Untersuchungs-
personen bei - 16ff

Beobachtungseinheit 11ff

Bewertung von - 22ff

-, Kategoriensystem 12ff

Planung von - 6ff

Reliabilität von - 22ff

Reproduzierbarkeit von - 22f

teilnehmende - 4, 5f

Validität von - 22f, 48ff

wissenschaftliche(n) - 1ff

Arten von - 3ff

Bias (Interviewer-Bias) 330

Biographische(n) Methode

(Biographik) 362ff

- in der Entwicklungs-
psychologie 365f

Forderungen an - 376

- in der humanistischen
Psychologie 367f

Idiographische Persön-
lichkeitspsychologie und - 364f

Objektivität der - 375ff

-, Probleme in der
Psychoanalyse 368ff

psychologische Streßforschung und
- 373ff

- in der Sozialisationsforschung 371ff
statistische Auswertung 386f

Butler & Haigh-Q-Sort 135ff

California Q-Set (CQ-Set) 136

Denken, lautes 91

Einzelfallanalyse 387ff

-, Datensammlung und Ver-
suchsplanung 389ff

-, statistische Auswer-
tungsmethoden 391ff

EPA-Struktur 159f

Ereignisstichprobe 20f

Erkundungsgespräch s. Exploration

Erlebnisbeschreibung 77f

Experiment 9, 326

Exploration 321ff, 383

Auswertung einer - 334f

- in der Eignungsdiagnostik 349ff
Geschichte der - 323f

methodische Prinzipien der - 329ff

-, qualitative Charakterisierung 325ff

-, Reliabilität 339f

-, Validität 345f

Exploratives Gespräch s. Exploration

Explorative Methode s. Exploration

- Facettentheorie 6ff, 15
- Feldbeobachtung 9
- Frage(n) 241
 - bedeutungsäquivalente - 261f
 - heikle - 274f, 277f
 - offene, geschlossene
 - Reihenfolge 267f, 270, 273f
- Fragebogen 222ff, 327
 - , Antwortkategorien 254f
 - , Antwortmotivation 269ff
 - Antwortverhalten in - 229f, 231f
 - Anwendungsgebiete von - 227f
 - äußere Gestaltung 279ff
 - Beantwortungsfehler in - 235
 - , Fragenformulierung 251ff, 261ff
 - , Frageninhalte 243f, 252ff
 - , Fragentypen 241, 246
 - , Gedächtnisprozesse 236f
 - Konstruktionsgesichtspunkte
 - von - 224f
 - , kontexteffekte 270ff
 - Systematik von - 222f
 - Verfälschbarkeit von - 229f
- Fremdbeobachtung 77ff
 - Verlässlichkeit der - 118
- Generalisierbarkeit(stheorie) 8, 41f
- Gespräch 321
- Halo-Effekt 104
- Idiographische Persönlichkeitspsychologie 364f
- Interview 322, 327
 - halbstrukturiertes - 383
- Introspektion 78ff
- ipsative Daten 135
- Kappa (x) von Cohen 30ff
- Kategoriensystem 12f
- Korrelation(skoeffizienten) (s.a. Übereinstimmungsmaße)
 - Intraklassen - 37ff
 - Produkt-Moment - 25, 37, 38
 - Rang - 34, 36
- Metakognition 97
- multitrait-multimethod matrix (MTMM) 51f
- Personen-Stichprobe 20
- Pi (n) von Scott 30
- Psychobiographie 370
- Q-Korrelation 135
- Q-Sort(-Verfahren) 135ff
 - , Anwendung 140f
 - , Auswertung 144
 - Butler & Haigh - 136ff, 142
 - , California Q-Set (CQ-Set) 136, 147
 - , Einordnung 135
 - , Itemselektion und Itemorganisation 141f
 - , Reliabilität 144f
 - , Validität 145f
- R-Daten 147
- Reliabilität 22, 23f, 46f
 - , pfadanalytische Modelle 42ff
 - varianzanalytische Ansätze 37ff
- Reproduzierbarkeit von (wissenschaftl.) Beobachtungen 22f
- Rho (q) von Spearman 34, 36
- S-Daten 147
- Schätzskalensystem 13
- Selbstbeobachtung 77ff
 - methodische - 83ff
- Verlässlichkeit der - 118
- Semantisches Differential 154ff
 - , Auswahl der Skalen 189ff
 - , Daten-Reduktionstechniken 161ff
 - , EPA-Struktur 151f
 - , interindividuelle Unterschiede 169ff
 - , Konzept-Skaleninteraktion 172ff
 - , Konzeptvarianz 162ff
 - , Reliabilität 181ff
 - , Skalenqualität 184ff
 - , systematische Urteilsfehler 178ff

- , transkulturelle Stabilität 168f
- zugrundeliegende Modelle 154ff
- Signalerkennungstheorie 107
- Übereinstimmung(smaße) 26ff
 - für Intervalldaten 37ff
 - , pfadanalytische Modelle 42ff
 - varianzanalytische Ansätze
 - , (Intraklassenkoeffizienten) 37ff
 - für Nominaldaten 29ff
 - , Π (π) 30
 - , G-Index 33
 - Kappa (κ) 30ff
 - , für Ordinaldaten 36f
 - , Gamma (γ) 36
 - , Rho (ρ) 34, 36
 - , W von Kendall 36
- prozentuale - (%Ü) 26ff
- Validität 22, 48ff
 - Konstrukt- 50ff
 - konvergierende und diskriminierende
 - 8, 51ff
 - Kriteriums- 49
- Verfälschbarkeit (von Fragebogen) 229f
- Verhaltensbeobachtung 99
- Verhaltenseinschätzung 117
- Vorstellungen, bildhafte 98f
- W-Konkordanzmaß von Kendall 36
- Zeichensystem 13
- Zeitstichprobe 18
- Zusammenhang 29

Autoren-Register

Hinweis: Die kursivgedruckten Seitenangaben beziehen sich auf die Literaturverzeichnisse der Artikel.

- Ach, N. 84ff, 88, 124
 Adair, J. 98, 124
 Adams, F. M. 200
 Adams, J. S. 260, 291
 Adams, R. S. 14, 60
 Adler, A. 370, 378
 Aheshwari, A. 207
 Aiken, E. G. 200
 Aivano, S. L. 62
 Ajzen, I. 10, 64
 Akuto, H. 216
 Alexander, S. 199, 200
 Algina, J. 387, 400
 Allison, R. B. 200
 Althausen, R. P. 57, 60
 Allport, G. W. 364, 371, 394
 Alutto, J. A. 287, 291
 Alwin, D. F. 57, 60, 75
 Amerine, M. A. 36, 60
 Amsfeld, Elizabeth 200
 Amster, Harriet 200
 Anastasi, A. 224, 229, 291
 Andersen, E. B. 231, 291
 Anderson, A. B. 168, 171, 185, 188, 200
 Anderson, R. L. 47, 71
 Anger, H. 241, 245, 248, 251f, 253, 257, 261, 264, 269f, 291, 308, 318, 322, 325, 356
 Angst, J. 292, 307, 319
 Anisfeld, M. 200
 Anton, J. L. 388, 394
 Antons, K. 356
 Arnold, J. B. 168, 201
 Arnold, W. 356
 Armstrong, J. S. 307, 318
 Arntzen, F. 304, 318
 Arrington, R. E. 18, 61
 Arthur, A. Z. 178, 179, 201
 Asendorpf, J. 26, 32, 37, 61
 Ash, R. 240, 289, 291
 Ashley, R. K. 271, 278, 301
 Ashton, P. T. 388, 394
 Assakul, K. 47, 61
 Assessment Staff 324, 356
 Atteslander, P. 223, 241, 291, 305, 318
 Atwood, J. T. 186, 201
 Avison, W. R. 57, 61
 Azrin, N. H. 123, 124
 Baade, W. 356
 Babbie, E. R. 306, 318
 Baton, F. 83, 124
 Baer, D. M. 122, 124, 128, 387, 392, 394, 398
 Bakan, D. 90, 124
 Bahrdrdt, H. P. 372, 394
 Bailey, K. D. 305, 310, 315, 318
 Bain, R. K. 332, 356
 Baker, F. B. 45, 54, 67
 Bales, R. F. 11, 13, 61
 Ballweg, J. A. 173, 177, 210
 Baltes, P. B. 305, 318
 Bannister, D. 392, 394
 Barclay, A. 201
 Barker, R. G. 377, 394
 Barkus, Ph. 334, 359
 Barker, R. 20, 62, 100, 124
 Barlow, D. H. 148, 151, 364, 388, 389, 394, 396
 Barnard, J. W. 201
 Barnard, P. 263, 281, 301
 Barnes, J. 286f, 296
 Barton, A. H. 245, 247, 291, 297
 Barrett, G. V. 201
 Bartenwerfer, H. 372, 394
 Bartko, J. J. 37, 61
 Bass, A. R. 210
 Baumann, D. J. 144, 148
 Baumann, U. 199, 201
 Baumgartner, R. 307, 319
 Baxter, J. C. 201
 Bean, A. G. 212
 Beck, R. C. 202
 Beck, W. 324, 356
 Behrens, K. C. 291
 Bellak, I. 368, 394
 Bellamy, G. T. 125, 128
 Bellows, R. 324, 356
 Belschner, W. 317, 318
 Belson, W. A. 256, 291
 Bern, D.J. 93, 125
 Benel, Denise, C. R. 201
 Benel, R. A. 201
 Bennett, L. F. 151
 Bente, G. 136, 151

- Bentler, P. M. 185, 199, 201
 Berdie, D. R. 278, 291, 307, 319
 Berg, I. A. 229, 292
 Bergan, J. R. 35, 61
 Bergermeier 104, 125
 Bergler, R. 154, 192, 201
 Berlyne, D. E. 201
 Berk, R. A. 288, 292
 Berman, J. S. 105f, 125
 Bernard, J. L. 204
 Bettinghaus, E. P. 201
 Bien, W. 57, 61
 Binder, J. 286f, 292, 307, 319
 Bingham, W. 333, 356
 Binswanger, L. 394
 Birch, D. 201
 Birdsall, T. G. 125, 131
 Birdwhistell, R. L. 100, 125
 Bishop, F. M. 334, 357
 Bishop, Y. M. 29, 61
 Black, H. K. 201
 Black, J. L. 127, 130
 Blakeney, R. N. 329, 356
 Blalock, H. M. Jr. 43, 44, 61
 Block, J. 125, 136, 140, 141, 142, 143, 144, 145, 147, 148, 149, 201, 229, 292
 Blumenthal, A. L. 82, 125
 Bobbitt, R. G. 202
 Bock, R. D. 17, 61
 Böcker, W. 334, 356
 Böttgen, F. 16, 61
 Böttger, P. 397
 Bogo, N. 200
 Bohrstedt, G. W. 43, 61
 Bois, J. 71, 131
 Bokander, I. 202
 Bolstad, O. D. 121, 123, 125, 128
 Bolton, D. L. 338, 356
 Boltin, N. 207
 Borden, R. 295
 Borg, I. 6, 7, 61, 104, 125
 Borgatta, E. F. 102, 125
 Boring 91, 125
 Bortz, J. 192, 202, 206
 Boruch, R. F. 56, 58, 61
 Bousfield, W. A. 202
 Boyer, E. G. 13, 73
 Braaten, K. 297
 Bradburn, N. M. 229, 247, 265, 267, 270, 272f, 278, 283f, 284, 292, 301, 307, 320
 Bradley, P. 293
 Brandt, L. W. 187, 202
 Brannigan, C. R. 62
 Braun, P. 90, 125
 Bredenkamp, J. 16, 47, 62
 Brennan, R. L. 32, 33, 62
 Brewer, W. F. 202
 Brinton, J. E. 202
 Brod, Diane 178, 202
 Broden, M. 122, 125
 Brown, J. D. 212
 Brown, R. 191, 202
 Bross, I. 47, 62
 Broughton, A. 140, 153
 Brown, A. L. 97, 125
 Bruner, J. S. 104, 125
 Brunswik, E. 110, 125
 Buchner, D. 285, 287, 292
 Bühler, C. 365, 366, 381, 394
 Bühler, K. 348, 356
 Burger, G. K. 202
 Burgess, E. W. 375, 395
 Burisch, M. 225, 226, 241, 292
 Burland, S. 130
 Burnhill, P. 295
 Burns, E. 145, 149
 Burns, R. 173, 202
 Butler, J. M. 136, 142, 146, 147, 149
 Butler, L. 98, 130
 Butler, R. P. 284, 292
 Bynner, J. 173, 177, 202
 Cahalan, D. 249, 292
 Calder, P. H. 320
 Caldwell, B. M. 13, 62
 Callahan, D. M. 320
 Cameron, R. 130
 Campbell, D. T. 16, 51, 52, 53, 54, 55, 56, 57, 58, 60, 62, 74, 125, 133, 267, 292, 380, 389, 395
 Cannell, C. F. 227, 236, 238, 241, 243, 245f, 247, 252, 253, 256, 263, 265, 268, 270, 272, 276, 287, 292, 333, 357
 Canter, D. 7, 62
 Cantril, H. 254, 257, 266, 271, 299
 Caplan, S. W. 140, 149
 Carl, W. 229, 292, 293
 Carpenter, E. H. 307, 319
 Carroll, J. B. 163, 174, 202
 Carroll, R. M. 202
 Carroll, S. J. 247, 293
 Carter, D. J. 185, 191, 220
 Carter, R. F. 202
 Cartwright, D. S. 144, 149
 Cartwright, R. D. 140, 149
 Casey, M. 293
 Cash, W. B. 241, 300
 Cassel, R. N. 202
 Castellan, N. J. Jr. 111, 125
 Cataldo, E. F. 277, 292
 Cattell, R. B. 105, 125, 135, 145, 147, 149, 225, 292, 299, 325, 356, 392, 395

- Cavanaugh, J. C. 98f, 125
 Cavior, N. 119, 122, 125
 Centra, J. A. 53, 62
 Chaffee, S. H. 202
 Champneys, F. H. 365, 395
 Chapman, J. P. 105, 126
 Chen, Kathleen 178, 202
 Chapman, L. J. 105, 125
 Chase, P. H. 140, 149
 Chetwynd, S. J. 73
 Chi, P. L. 107, 126
 Chiba, Y. 215
 Child, I. L. 75
 Chipman, S. 126, 132
 Chmura, Kathy J. 212
 Christensen, A. 126, 128
 Chicchetti, D. V. 26, 62
 Ciminero, A. R. 117, 121, 126, 129
 Clanton, E. S. 296
 Claparede, E. 91, 126
 Clark, H. H. 203
 Clark, M. L. 203
 Clark, S. 184, 186, 220
 Clark, V. A. 186, 203
 Clarke, A. H. 2, 62
 Clauss, G. 249, 292
 Cleary, T. A. 48, 62
 Clevenger, T. 203
 Cliff, N. 187, 203, 229, 292
 Coan, R. W. 225, 293
 Cochran, W. G. 16, 48, 62
 Cohen, J. 30, 31, 33, 47, 62, 65, 143, 144, 149, 151
 Cohen, R. 105, 126, 229, 293
 Cohen, S. H. 48, 67
 Comte, A. 79, 82, 126
 Conger, A. J. 31, 56, 62
 Cook, S. W. 67, 180, 221
 Coombs, C. H. 107, 109, 126
 Corby, G. 299
 Corsini, R. J. 140, 149
 Cosper, R. 356
 Costner, H. L. 43, 44, 57, 63
 Cottrell, L. S. 125
 Cowen, L. 203
 Coyle, B. W. 71
 Coyle, F. A. 144, 149
 Coyne, L. 182, 203
 Craighead, W. E. 126, 128
 Crampton, C. W. 346, 356
 Cranach, M. v. 2, 4, 11, 12, 13, 63, 68
 Crapsi, L. A. 212
 Creelman, M. B. 203
 Crockett, W. H. 171, 203
 Cronbach, L.J. 41, 63, 65, 104, 126, 144, 145, 149, 188, 203, 224, 226, 229, 231, 289, 293
 Crutschfield, R. S. 293
 Cureton, E. E. 36, 63
 Dahme, B. 392, 395
 Dailey, C. A. 375, 395
 Damarin, F. 233, 293
 D'Andrade, R. G. 102f, 104, 108, 126
 Danziger, K. 91, 126
 Darnell, D. K. 172, 186, 190, 203
 Das Rhea, S. 167, 210
 Darwin., C. 365, 395
 Davis, A. 371, 395
 Davis F. B. 49, 63
 Dawes, R. M. 126, 188, 203
 De Boor, C. 370, 395
 De Burger, R. A. 203
 Deese, J. 184, 203
 Deming, W. E. 16, 63
 Denmark, F. L. 170, 203
 Deo, P. 146, 149
 Deters, H. 352, 356
 Deutsch, M. 67
 Deutschmann, P. J. 203
 De Waele, J. P. 375, 395
 Diab, L. N. 203
 Diament, C. 71, 126, 132
 Dice, L. R. 35, 63
 Dick, W. 181, 204
 Dicken, C. F. 203
 Dickman, H. R. 63
 Dickson, J. P. 283, 285, 293
 Dickson, W. J. 324, 359
 Diehl, B. 184, 188, 204
 Dietz, A. 127
 Dijkstra, W. 265, 297
 Dillehay, R. 258, 290, 293
 Dilthey, W. 362, 395
 Dilts, M. 204
 Dimock, H. S. 346, 357
 Dittrich, A. 199, 201
 DiVesta, F. 181, 204
 Dörner, D. 91, 126
 Dohrenwend, B. S. 299, 333, 359, 384, 399
 Dollard, J. 371, 375, 377, 395
 Donahoe, J. W. 203
 Dorsch, F. 135, 149
 Doyle, K. C. 290, 293
 Dudycha, A. L. 71
 Dukes, W. F. 364, 395
 Duncan, O. D. 43, 63
 Duncan, S. 100, 126
 Duncan, S. Jr. 2, 14, 64
 Duncker, K. 91, 226
 Dyke v. & Moore, B. V. 356
 Dymond, R. F. 140, 144, 145, 146, 147, 149, 152
 Ebel, R. L. 37, 38, 64
 Eber, W. H. 292
 Echelmeyer, L. 325, 357
 Edgell, S. L. 240, 289, 291
 Edinger, L. J. 370, 395

- Edwards, A. L. 224, 228f, 251, 293
 Egan, J. P. 107, 109, 127
 Ehlers, T. 230, 242, 261, 293
 Eisenman, R. 204
 Eiser, J. R. 180, 204
 Ekman, G. 168f, 211, 336, 357
 Ekehammar, B. 148, 150
 El-Batrawi, S. 144, 152
 Ellgring, J. H. 2, 62
 Elliott, L. L. 199, 204
 Ellis, A. 245, 293
 Endicott, J. 65
 Endler, N. S. 204
 Ends, E. J. 147, 150
 Engel, M. 150
 Engquist, G. 11, 71, 112, 127, 131
 Erbslöh, E. 310, 319, 322, 357
 Epperson, W. V. 284, 293
 Erdos, P. L. 242, 247, 262, 279f, 281, 283, 287, 293
 Erickson, M. 201
 Ericsson, K. A. 91, 114, 116, 127
 Erikson, H. E. 370, 395
 Ertel, S. 164, 167, 171, 172, 173, 185, 189, 204
 Ervin-Tripp, S. M. 205
 Esser, H. 312, 319
 Estep, M. F. 324, 356
 Etkind, A. M. 205
 Everitt, B. S. 32, 64, 65
 Everett, A. V. 168, 205
 Eysenck, H. J. 226f, 227, 246, 259, 293, 353, 357
 Fagan, J. 147, 150
 Falkenberg, S. D. 186, 201
 Falthzik, A. M. 247, 293
 Farr, S. D. 173, 210
 Faßnacht, G. 2, 11, 64
 Fear, R. A. 334, 357
 Feenstra, H. J. 206
 Feger, B. 88, 127
 Feger, H. 1, 2, 10, 11, 13, 18, 42, 43, 53, 57, 64, 71, 88, 113, 127, 132, 193, 205, 224, 294, 388, 395
 Feldt, L. S. 290, 301
 Fertig, J. 184, 219
 Fiedler, F. E. 140, 150
 Field, J. 202
 Fienberg, S. 61
 Filipp, S. H. 148, 150
 Finke, H. O. 295
 Finley, J. R. 205
 Finn, R. H. 40, 64
 Finke, R. A. 95, 127
 Fischer, G. 227, 231, 266, 287, 294
 Fischer, G. W. 48, 64
 Fischer, G. H. 148, 150
 Fischer, K. 354, 360
 Fishbein, M. 10, 64, 170, 171, 172, 176, 217, 220
 Fiske, D. W. 14, 51, 52, 53, 54, 55, 56, 57, 58, 60, 63, 64, 229, 267, 292, 294, 301
 Fisseni, H. J. 289, 294, 319, 347, 357
 Fixen, D. L. 119, 127
 Flade, A. 192, 205
 Flavell, J. H. 98, 127, 205
 Fleiss, J. L. 32, 33, 34, 35, 38, 46, 47, 64, 341, 357
 Fode, K. 206
 Fodor, J. A. 205
 Förster, P. 289, 294
 Fang, C. 36, 73
 Ford, L. H. 180, 205, 212
 Forthman, J. H. 198, 205
 Franke, J. 192, 206
 French, G. M. 185, 211
 Friedman, C. J. 206
 Foster, D. F. 399
 Foster, S. L. 68, 119f, 123f, 127, 129
 Fowler, R. D. 144, 149
 Fox, W. C. 127, 131
 Frank, G. H. 144, 146, 150
 Frawley, W. H. 36, 72, 72
 Freeman, W. 71
 Frenz, H. G. 2, 4, 11, 13, 65
 Freud, A. 370
 Freud, S. 368, 371, 378, 395
 Frey, D. 310, 319
 Frey, S. 11
 Frick, R. 26, 65
 Friedman, I. 140, 147, 150
 Friedrich, W. 241, 244f, 286, 289, 294
 Friedrichs, H. 325, 357
 Friedrichs, J. 5, 65, 227, 241, 245ff, 252, 254, 256, 263, 271, 294, 302, 319, 379, 395
 Frisbie, B. 248, 294
 Froehlich, C. P. 333, 357
 Froehlich, R. E. 357
 Frohburg, I. 136, 137, 138, 139, 144, 146, 147, 150
 Fuchs, A. 171, 183, 187, 188, 193, 194, 206
 Fürntratt, E. 229, 294
 Fuller, C. 283, 294, 307, 319
 Futrell, C. M. 284, 294
 Gärling, T. 168, 206
 Galanter, E. 14, 70
 Galtung, J. 255, 294

- Gardiner, D. H. 211, 221
 Gardner, R. C. 199, 206, 218
 Garfield, S. L. 140, 150
 Garrison, K. R. 372, 396
 Gebhard, P. H. 357
 Gehlen, H. 228, 297
 Gerbasi, K. C. 229, 231, 300
 Gesemann, G. 363, 396
 Getzels, J. W. 233, 294
 Gibson, J. J. 93, 127
 Giesen, H. 372, 394
 Giles, H. 211
 Gilpin, A. R. 186, 206
 Ginsberg, G. P. 297
 Giorgi, A. 99, 127
 Girard, R. 293
 Glad, B. 370, 396
 Gladden, J. W. 206
 Glasgow, W. D. 90, 127, 131
 Gleser, G. C. 41, 42, 63, 61, 145, 149, 183, 188, 203, 293
 Godfrey, R. R. 166, 206
 Goldberg, R. 295
 Goldberger, A. S. 43, 66
 Golden, G. H. 211
 Goldfried, M. R. 121, 127, 185, 197, 206, 207
 Goldiamond, I. 127
 Golding, S. L. 56, 58, 65
 Goldman, R. D. 212
 Goldstein, H. 396
 Goode, W. J. 269f, 273, 275, 282, 285f, 294
 Gooding, C. T. 144, 253
 Goodman, L. A. 29, 36, 65
 Gordon, D. A. 293
 Gottfried 117, 127
 Gottmann, J. M. 392, 396
 Gorospe, F. H. 206
 Graham, J. R. 140, 150
 Graumann, C. F. 1, 2, 3, 10, 11, 18, 65, 78, 82f, 127
 Gray, A. 281, 294
 Green, D. M. 28, 65
 Green, P. E. 168, 185, 197, 207
 Green, R. F. 207
 Greeno, J. G. 101, 128
 Griesemer, H. A. 288, 292
 Grigg, A. E. 207
 Grimm, G. 178, 191, 207
 Groeben, N. 317, 320
 Grooms, R. R. 215
 Gross, H. 324, 357
 Grubbs, F. E. 39, 65
 Grüner, K. W. 2, 5, 11, 13, 65, 128
 Gruhle, H. W. 396
 Gruson, L. 130
 Guertin, W. H. 142, 150
 Guilford, J. P. 66, 101, 106, 128, 145, 147, 150, 226, 249, 294
 Gullahorn, J. E. 280, 294
 Gullahorn, J. T. 280, 294
 Gulliksen, H. 109, 128, 182, 198, 207, 227, 294
 Gurwitsch, A. 92, 128
 Guski, R. 288, 295
 Gustad, J. W. 333, 356
 Guthrie, G. M. 140, 147, 150, 152
 Guttman, L. 6, 7, 8, 9, 66, 69, 266, 301
 Guttman, R. 6, 7, 8, 66
 Haan, N. 150
 Haase, H. 279, 282, 295
 Häcker, H. 230, 295
 Hagen, R. L. 121, 128
 Haggard, E. A. 37, 38, 66
 Haigh, G. V. 136, 142, 146, 149
 Haith, M. M. 365, 396
 Hall, C. 371
 Hall, R. V. 125i
 Hambitzer, M. 373, 396
 Harnikon, D. L. 178, 186, 207
 Hamilton, H. W. 207
 Hampel, R. 230, 295
 Hanlon, R. E. 147, 150
 Hannan, M. T. 43, 66
 Hannon, J. E. 204
 Hansen, M. H. 16, 66
 Harman, H. H. 167, 207
 Hart, J. T. 98, 228
 Hartemink, B. G. 359
 Hartley, J. 279, 281, 295
 Hartley, M. W. 140, 150
 Hartmann, H. 251, 295, 370
 Hase, H. 295
 Hasemann, K. 2, 66
 Hastorf, A. H. 207
 Hathaway, S. R. 242, 276, 295
 Hatt, P. K. 269f, 273, 275, 282, 285f, 294
 Hauser, R. M. 43, 66
 Haycock, V. 199, 220
 Hayes, D. P. 11, 66, 277f, 295
 Haygood, R. C. 184, 207, 218
 Heaps, R. A. 207
 Heard, W. G. 218
 Heberlein, T. A. 57, 60, 307, 319
 Heeler, R. M. 56, 71
 Heise, D. R. 43, 56, 66, 154, 167, 175, 188, 195, 199, 207
 Heller, D. 250, 295
 Hendel, D. D. 24, 66
 Henderson, E. H. 211
 Hendrick, C. 287, 295
 Hendrix, L. 290, 295
 Hennes, J. D. 220

- Hennig, W. 241, 271f,
294, 295
 Hensley, W. E. 319
 Herbert, E. W. 122, 128
 Herder, J. G. 362, 396
 Herrmann, T. 178, 207,
250, 295
 Hersen, M. 148, 151,
364, 388, 389, 394, 396
 Heskin, K. J. 173, 207
 Heyns, R. W. 2, 66
 Hickey, M. E. 338, 356
 Higa, M. 184, 210
 Hildebrand, D. K. 36,
66
 Hilden, A. H. 151
 Hill, R. J. 273, 301
 Hirschberg & Jones,
L.E. 208
 Hoar, J. R. 207
 Hörmann, H. 154, 207
 Hoeth, F. 230, 242, 295
 Hoffmann, K. 208
 Hofstätter, P. R. 147,
150, 190, 192, 208
 Hogg, J. 208
 Holeway, R. E. 197, 219
 Holland, P. 61
 Hollenbeck, A. R. 26,
27, 28, 32, 45, 66
 Holley, J. W. 33, 66
 Holm, K. 226, 234, 241,
243f, 249, 251, 264,
267, 270, 272, 276, 295,
296, 298
 Holtzmann, W. H. 396
 Holzkamp, K. 128
 Holzman, P. S. 182, 203
 Holz, W. 124
 Homzie, M. J. 208
 Hopper, D. 208
 Horman, M. 208
 Horn, J. L. 208
 Horner, A. J. 396
 Hornick, C. W. 225,
296
 Hornke, L. 290, 296
 Hornstein, H. A. 320
 Horowitz, M. 71
 Hovland, C. I. 180, 208
 Howe, E. S. 179, 187,
208
 Huber, H. P. 48, 66,
392, 396
 Hubert, L.J. 32, 33, 36,
44, 45, 54, 67
 Huck, S. W. 67
 Hübner, P. 298
 Humphries, D. A. 62
 Hundal, B. S. 146, 249
 Hunter, J. E. 48, 67
 Hunt, R. G. 212
 Hurley, J. R. 140, 152
 Hutwitz, D. 199, 208
 Hutwitz, W. N. 67
 Husek, T. R. 199, 200,
208
 Husen, T. 339, 357
 Hutt, C. 2, 18, 67
 Hutt, S. J. 2, 18, 67
 Inbau, F. E. 332, 333,
357
 Ingleby, J. D. 108f, 128
 Irish, D. P. 297
 Irle, M. 181, 208
 Irwin, L. 71
 Jackson, D. N. 55, 56,
58, 67, 105, 129, 180,
219
 Jacobs, P. I. 128, 133
 Jacobson, A. L. 43, 44,
67
 Jaehner, D. 377, 396
 Jahoda, M. 2, 67
 Jakobowitz, L. A. 199,
209, 211
 James, C. 212
 James, L. R. 296
 James, V. A. 324, 360
 James, W. 79, 81, 128
 Janke, W. 230, 287, 296,
312, 319
 Jannsen, J. P. 230, 296
 Janson, S. 35, 67
 Jaspers, D. 375, 377,
387, 396
 Jeanneret, P. R. 298
 Jenkins, E. 145, 149
 Jenkins, J. J. 209
 Jenkins, W. S. 329, 361
 Jernigan, L. R. 258, 290,
293
 Jetzschmann, H. 241,
296
 Joe, G. W. 67
 Jöreskog, K. G. 42, 43,
56, 67, 68, 71, 74, 75,
133, 304, 319
 John, D. 151
 Johnsgard, K. W. 217
 Johnson, A. 179, 205
 Johnson, C. A. 206
 Johnson, D. M. 106,
120ff, 128
 Johnson, J. H. 308, 319
 Johnson, R. L. 209
 Johnson, R. M. 274, 292
 Johnson, S. M. 106,
120ff, 128
 Johnson, W. R. 274, 296
 Jones, A. 143, 151, 296
 Jones, E. 369, 396
 Jones, J. M. 209
 Jones, R. R. 118, 129
 Jonsson, C. O. 241, 296
 Jordan, J. E. 15, 68
 Jung, C. G. 329, 357,
370
 Kahn, R. L. 241, 245,
253, 263, 272, 292, 296,
333, 357
 Kahneman, D. 172, 174,
178, 209
 Kalbermatten, U. 12, 68
 Kalinowsky-Czech,
M. 229, 296
 Kalis, B. L. 151
 Kallabis, H. 241

- Kalleberg, A. L. 56, 68
 Kamp, K. 379, 395
 Kane, A. 146, 151
 Kane, R. B. 196, 209, 277, 296
 Kanfer, F. H. 90, 122, 129
 Kanowitz, J. 131
 Kant, I. 77, 79, 82, 129
 Kanungo, R. N. 209, 212
 Kaplan, K. J. 186, 198, 209
 Kaplitza, G. 241, 255, 260, 267, 278, 280, 296
 Kardiner, A. 371, 377, 396
 Karmasin, F. 241, 244f, 248, 252, 254, 257, 259f, 261f, 264, 266f, 270f, 273f, 278, 281, 287, 296
 Karmasin, H. 241, 244f, 248, 252, 254, 257, 259f, 266f, 270f, 273f, 278, 281, 287, 296
 Kashiwagi, S. 209
 Katkin, E. S. 105, 133
 Katz, D. 92, 129
 Katz, M. 209
 Kaufman, H. J. 209
 Kaufman, I. C. 13, 68
 Kavanagh, M. J. 60, 68
 Kaye, K. 35, 46, 68
 Kazdin, A. E. 122, 129, 364, 396
 Keehn, J. D. 215
 Keeser, W. 392, 396, 399
 Keil, A. 209
 Keil, C. 209
 Keil, W. 151, 230, 261, 296
 Kellstedt, L. A. 292
 Kelly, G. A. 142, 151
 Kelly, J. A. 210
 Kemmler, L. 325, 357
 Kemnitzer, L. S. 140, 151
 Kendall, M. G. 36, 68
 Kennedy, D. R. 141, 152
 Kenny, D. A. 68, 105f, 125
 Kent, R. N. 2, 68, 71, 117, 119f, 123f, 127, 129, 131f
 Kentler, H. 210
 Kepes, S. Y. 284, 296
 Kephart, N. 324, 357
 Kerkhoff, T. 17, 71
 Kerlinger, F. N. 151
 Kernoff, P. 202
 Kerrick, J. S. 203
 Kessen, W. 365, 396
 Keys, A. 46, 68
 Kihlberg, J. K. 46, 68
 Kilty, K. M. 210
 King-Fun Li, A. 210
 Kinsey, A. C. 257, 296, 323, 331, 332, 333, 334, 335, 343, 344, 346, 357, 384, 396
 Kipnowski, A. 397
 Kirby, D. M. 206
 Kirchner, F. Th. 380, 397
 Kiresuk, T. J. 392, 397
 Kirschhofer-Bozenhardt, A. v. 241, 258, 260, 267, 278, 280, 296
 Kish, G. B. 286f, 296
 Kish, L. 16, 68
 Kissel, E. 397
 Kitay, D. H. 369, 397
 Kjeldergaard, P. M. 184, 210
 Klages, L. 340, 357
 Klapper, J. T. 210
 Klauer, K. J. 49, 68, 141, 148, 151
 Klebelsberg, D. v. 328, 357
 Klein, D. 299, 333, 359, 384, 399
 Klemmack, D. L. 173, 177, 210
 Klinkhammer, F. 230, 291
 Kluckhohn, F. R. 5, 68
 Kluegel, J. R. 56, 68
 Knudsen, D. D. 283, 297
 Koch, G. G. 26, 35, 37, 47, 68, 69
 Köbler, V. 242, 295
 Köhler, W. 93, 129
 Koen, F. 210
 König, R. 2, 68, 241, 297
 Kogan, N. 133
 Kohli, M. 362, 372, 397
 Komorita, S. S. 210
 Koolwijk, J. v. 235, 240f, 274, 297, 303, 319
 Koomen, W. 265, 297
 Korman, M. 210
 Korner, A. F. 37, 42, 69
 Kornreich, B. 146, 151
 Kosslyn, S. M. 95, 127
 Kostić, D. 167, 210
 Kraemer, H. C. 37, 42, 69
 Kratochwill, T. R. 364, 388, 397
 Krause, I. 210
 Krause, M. S. 54, 69
 Krauss, H. 50, 74
 Kraut, A. I. 273f, 278, 297
 Kravetz, S. 219
 Kreipe, K. 324, 358
 Kreutz, H. 241f, 246, 251, 255, 257ff, 262ff, 267, 269f, 274f, 286, 297, 313, 319
 Krieger, M. H. 172, 180, 210
 Krippendorff, K. 17, 35, 38, 40, 69
 Kris, E. 370
 Kriz, J. 30, 69
 Kröber, W. 324, 358
 Kroes, W. H. 184, 210
 Krüger, P. 250, 295

- Kruglanski, A. W. 16, 69
 Kruskal, W. H. 29, 36, 65, 69
 Kubinieć, C. M. 173, 210
 Kuiper, P. C. 363, 370, 397
 Kumata, H. 168, 210
 Kuncel, R. B. 229, 297
 Kunkel, E. 334, 335, 336, 348, 358
 Kuusinen, J. 164f, 176, 199, 211

 Lachman, J. L. 99, 129
 Lachman, R. 99, 129
 Laing, J. D. 66
 Lalu, N. M. 43, 44, 67
 Lambert, W. E. 200
 Lambert, W. W. 75, 100, 129
 Lampl de Groot 370, 397
 Lana, R. E. 211
 Landis, D. 219
 Landis, J. R. 26, 35, 37, 65, 69
 Lane, S. T. M. 211
 Langer, I. 135, 151
 Lansing, J. B. 227, 236, 297
 Lantermann, E. D. 297
 Larkin, J. D. 62
 Laurent, A. 292
 LaVoie, A. L. 199, 201
 Lawlis, G. F. 40, 69
 Lawson, E. D. 211
 Laxer, R. M. 141, 152
 Lay, C. H. 105, 129
 Lazarsfeld, P. F. 247, 252, 297
 Lazier, G. A. 203
 Leach, G. M. 69
 Ledvinka, J. 358
 Lehr, U. 334, 342, 348, 358, 372, 397

 Leitenberg, H. 364, 388, 389, 397
 Lennertz, E. 288, 297
 Levin, J. 171, 211
 Levy, L. H. 146, 147, 151, 177, 210
 Levy, N. 185, 216
 Levy, P. 9, 69, 211
 Levy, S. 42, 69
 Lewin, K. 329, 358
 Lewis, O. 371, 397
 Libby, W. L. 184, 210
 Lichtenstein, E. H. 202
 Lieberman, D. A. 90, 129
 Lienert, G. A. 26, 33, 66, 69, 227, 266, 287, 297, 336, 358
 Light, C. S. 178, 211
 Light, R. J. 30, 31, 32, 33, 62, 69
 Lilli, W. 181, 211
 Lindsey, D. 295
 Linehan, M. M. 121, 127
 Linn, R. L. 48, 56, 62, 71, 74, 75, 133
 Linschoten, J. 229
 Linsky, A. S. 279, 283, 285, 287, 297, 307, 319
 Linton, R. 375, 397
 Lipinski, D. P. 119, 126, 129f
 Lipmann, O. 356
 Lisch, R. 30, 69
 Litt, E. N. 171, 211
 Litwak, E. 257, 297
 Livson, N. H. 144, 151
 Lobitz, G. K. 120, 128
 Loch, W. 370, 297
 Loewenstein, R. M. 370, 398
 Lohr, J. M. 166, 211
 Long, B. 178, 179, 211
 Longabaugh, R. 2, 14, 17, 20, 23, 70, 99, 130
 Lopez, F. M. 338, 354, 358
 Lord, F. M. 43, 70

 Lorr, M. 105, 130
 Louiselle, R. M. 221
 Lowy, D. G. 221
 Lu, E. 40, 69
 Lu, K. H. 38, 70
 Luce, R. D. 14, 70
 Lück, H. E. 207
 Lüdtke, H. 5, 65
 Lüer, G. 91, 111, 130
 Luria, Z. 214
 Lysterly, S. B. 36, 70
 Lyle, J. 211

 Maccoby, E. E. 241, 244, 248, 251, 259, 271f, 298
 Maccoby, N. 241, 244, 248, 251, 298
 Mac Kinney, A. C. 62, 68
 Maclay, H. 211
 Madden, J. E. 211
 Madow, W. G. 66
 Magnussen, D. 227, 298
 Magnusson, D. 168, 211, 337, 358
 Magnusson, F. 168, 220
 Maguire, T. O. 211
 Makohoniuk, G. 118, 130
 Maletzky, B. M. 121, 130
 Malmstrom, E. J. 185, 211
 Maltz, H. E. 178, 212
 Manis, M. 212
 Mann, F. 276, 286, 298
 Mann, J. H. 125
 Mann, R. D. 102f, 130
 Manz, W. 2, 13, 70
 Marabotto, C. 119, 122, 126
 Marcill J. C. 147, 148, 151
 Margis, P. 323, 358
 Markel, N. N. 184, 212

- Marks, E. S. 243, 272, 298
 Marks, I. 178, 212
 Marks, P. A. 140, 144, 149, 151
 Marquis, K. H. 292
 Martin, C. E. 296, 333, 357, 396
 Martinez, J. L. jr. 212
 Martinez, S. R. 212
 Maruyama, K. 183, 212
 Mash, E. J. 118, 130
 Mash, L. J. 130
 Maslow, A. 367, 398
 Mason, W. M. 267, 270, 272f, 278, 292
 Massarik, F. 367, 381, 394
 Matteson, M. T. 307, 320
 Mauldin, W. P. 243, 272, 298
 Mauxois, A. 377
 Maxwell, A. E. 37, 70
 May, W. H. 154, 165, 176, 215, 219
 Mayer, L. S. 43, 33, 70
 Mayerberg, C. K. 212
 Mayfield, E. C. 342, 353, 358
 Mayntz, R. 241, 251, 262, 298
 Mazlish, B. 370, 398
 McCall, G. J. 5, 6, 70
 McCallon, E. L. 212
 McClave, J. T. 74
 McCormick, E. J. 289, 298
 McDermott, P. A. 31, 70, 74
 McDonagh, E. C. 286, 298
 McElwee, J. D. 118, 130
 McFarlane, J. W. 323, 358
 McGinnies, E. 216
 McGrew, W. C. 2, 70
 McGuire, W. J. 380, 386, 398
 McKelvie, S. J. 198, 212, 298
 McKinley, J. C. 242, 276, 295
 McKinley, S. M. 29, 70
 McNair, D. M. 130
 McNaughton, J. F. 329, 336
 McNicol, D. 28, 70, 107, 130
 Mecham, R. C. 298
 Medley, D. M. 2, 13, 14, 70
 Meehl, P. E. 63, 226, 293
 Meek, E. E. 207
 Mees, U. 2, 71
 Meichenbaum, D. 97f, 130
 Meier, R. D. 334, 358
 Meisels, M. 180, 205
 Mellenbergh, G. J. 109, 129, 133, 359
 Meltzer, L. 66
 Messer, S. 212
 Messick, S. J. 187, 188, 212
 Metge, A. 82f, 130
 Metzger, W. 92, 130, 334, 358
 Metzler, J. 95f, 132
 Metzner, H. 276, 286, 298
 Meurer, K. 144, 145, 152
 Michel, L. 358
 Micko, H. C. 192, 212
 Middleton, D. 107, 133
 Mierke, K. 324, 358, 384, 398
 Miettinen, O. S. 398
 Mikula, G. 178, 191, 192, 212
 Milbrath, L. W. 292
 Miles, M. B. 320
 Miller, F. D. 320
 Miller, G. A. 212
 Miller, P. McC. 212
 Miller, R. L. 310, 316, 320
 Miller, S. 212
 Mills, D. H. 192, 213
 Mindak, W. A. 197, 213
 Minsell, W.-R. 136, 151
 Miron, M. S. 154, 161f, 182, 190, 213, 215
 Misch, E. 363, 398
 Mitchell, D. B. 95, 130
 Mitsos, S. B. 178, 191, 192, 213
 Mittenecker, E. 224f, 231, 233, 242, 298
 Mitts, B. 125
 Mitzel, H. E. 2, 13, 14, 70
 Moffatt, G. W. 353, 359
 Mogar, R. E. 178, 213
 Moore, B. V. 333, 356
 Moos, R. H. 121, 130
 Mordkoff, A. M. 185, 213
 Morimoto, H. 213
 Morland, J. K. 220
 Morris, C. 215
 Moss, C. S. 213
 Mote, V. L. 71
 Mower White, C. J. 180, 204
 Mowrer, H. H. 135, 151
 Mucchielli, R. 241, 298
 Mueller, W. S. 213
 Münch, W. 241, 298
 Mulaik, A. 105, 130
 Munoz, S. R. 200
 Munroe, R. L. 20, 71
 Murray, E. J. 295
 Murray, H. A. 368, 377
 Mutherr, B. 75, 172, 213
 Nahinsky, I. D. 140, 151
 Nanda, H. 63, 293
 Narayana, C. L. 249, 298

- Natalicio, L. F. S. 166, 206
 Natsoulas, T. 90, 130
 Nay, W. R. 17, 71
 Naylor, J. C. 17, 71
 Nee, J. C. M. 65
 Neff, W. S. 143, 151
 Nelson, R. O. 119, 126, 129, 130
 Neufeld, R. W. 388, 398
 Neuringer, C. 178, 213
 Newcomb, T. M. 104, 131
 Newton, D. 11, 71, 112, 131
 Nicewander, W. A. 47, 71
 Nichols, H. J. 217
 Nichols, T. F. 144, 151
 Nickels, S. A. 179, 191, 213
 Nidorf, L. J. 171, 203
 Niederland, W. G. 369, 398
 Nisbett, R. E. 79, 131, 316, 320
 Nishimura, H. 216
 Noelle, E. 241f, 248, 250, 259f, 266f, 278, 298
 Noelle-Neumann, E. 239, 252ff, 262, 266, 268f, 271ff, 275, 278, 298
 Nordenstreng, K. 164f, 168, 172, 175, 213, 214
 Norman, W. T. 105, 131, 181, 214
 Notarius, C. 392, 396
 Novick, M. R. 43, 70
 Nowakowska, M. 229, 233, 298
 Nunnally, J. C. 146, 152
 Nußbaum, A. 42, 71
 Nuttin, J. 92, 131
 Obst, G. H. 372, 398
 O'Connell, E. J. 62
 O'Connor, J. 147, 150
 O'Donovan, D. 191, 214
 Oetting, E. R. 191, 214
 Olbrich, M. 349, 359, 398
 Oldfield, R. C. 324, 359
 O'Leary, K. D. 17, 124, 129, 131, 132
 Oles, H. J. 182, 214
 Olmedo, E. L. 212
 Olsson, U. 213
 O'Mally, J. M. 217
 Ono, H. 207
 Oppenheim, A. N. 241, 264, 298
 Orlik, P. 167, 178, 191, 214
 Orne, M. T. 333, 359
 Osgood, C. E. 154f, 172ff, 188ff, 200, 214, 215, 218, 219, 250, 261, 298
 Osipow, S. 215
 Osterland, M. 372, 398
 Otis, J. L. 201
 Overall, J. E. 39, 71
 Oyama, T. 214, 218
 Padden, D. 208
 Page, C. W. 147, 150
 Paivio, A. 215
 Paponia, N. 71
 Papousek, H. 2, 73
 Parkinson, C. N. 324, 359
 Parloff, M. B. 147, 148, 153
 Parouson, B. S. 392, 398
 Paskewitz, D. A. 333, 359
 Passini, F. T. 105, 131
 Pastore, R. E. 107, 131
 Patterson, G. R.
 Paul, G. L. 118, 128, 131
 Paul, S. 372, 398
 Pauleinkhoff, B. 398
 Pauli, R. 356
 Pauling, F. J. 211
 Payne, S. L. 247, 249, 251ff, 258, 261, 263f, 266, 282, 298
 Peabody, D. 178, 215
 Peak, H. 2, 71
 Pearson, K. 34, 71
 Peck, R. C. 284, 293
 Peckham, S. 201
 Perkins, H. V. 140, 152
 Perlmutter, M. 98f, 126
 Perreault, W. D. 269, 277, 279, 282, 290, 299
 Petermann, F. 364, 380, 381, 387, 388, 392, 397, 398, 399
 Petermann, U. 381, 399
 Peters, D. J. 73
 Peters, L. H. 260, 301
 Peterson, A. O. D. 152
 Peterson, D. R. 131
 Peterson, W. W. 104, 131
 Pettersson, T. 213
 Pfahler, G. 323, 359
 Phillips, B. S. 241, 254, 257, 259f, 268, 270, 272, 299
 Phillips, E. L. 127, 152
 Piaggio, L. 181, 215
 Pickett, L. 202
 Pilkington, C. W. 131
 Pilliner, A. E. G. 37, 70
 Plutchik, R. 215
 Podgorny, P. 131
 Polansky, N. 5, 71
 Pomeroy, J. E. 296
 Pomeroy, W. B. 333, 357, 396
 Pongratz, L. 323, 359
 Pool, J. 11, 65
 Pope, B. 208
 Pope, H. 297
 Postman, L. 110, 131
 Powell, G. E. 73
 Prager, R. A. 140, 150
 Presly, A. S. 173, 177, 215

- Presser, S. 312, 320
 Preyer, W. 365, 399
 Price, J. M. 47, 71
 Price, R. H. 107, 131
 Priest, P. N. 178, 215
 Procter, C. H. 47, 61
 Prothro, E. T. 215

 Quarter, J. 141, 152

 Raab, E. 263, 266, 299
 Rachlin, H. 122, 131
 Radford, J. 90, 131
 Raiford, A. 144, 152
 Rajaratnam, N. 38, 63, 72, 206, 293
 Rao, P. V. 74
 Rao, V. R. 207
 Rapaport, D. 71
 Rappaport, J. 204
 Rasmussen, V. 399
 Ray, M. L. 56, 72
 Ray, W. S. 152
 Rechetnick, J. 359
 Redfield, J. 118, 131
 Reece, M. W. 215
 Reed, T. R. 215
 Reid, J. B. 18, 72, 118f, 132
 Reid, J. E. 332, 333, 357
 Renzaglia, G. A. 120, 132
 Revenstorff, D. 167, 171, 174, 180, 187, 188, 199, 216, 392, 399
 Revie, V. A. 140, 152
 Richardson, J. T. E. 94, 132
 Richardson, S. A. 241, 299, 333, 359, 384, 399
 Richman, C. L. 95, 130
 Richter, H. J. 241, 243, 247, 249, 266, 269, 273, 276, 279ff, 284f, 286ff, 299
 Riley, R. T. 203

 Rindner, R. J. 112, 131
 Ring, E. 256, 281, 299
 Risley, T. R. 388, 394, 399
 Roberts, R. R. Jr. 120, 132
 Robinson, P. W. 399
 Robinson, R. 219
 Rock, D. A. 48, 72
 Roessler, E. B. 36, 60
 Roethlisberger, F. J. 324, 359
 Rogers, A. H. 140, 152
 Rogers, C. R. 140, 146, 147, 152
 Rogers, T. B. 229, 299
 Rogot, E. 47, 72
 Rohrachner, H. 82, 132
 Rohrmann, B. 229, 249, 264, 295, 299
 Roll, S. 216
 Romanczyk, R. G. 17, 18, 72, 118, 132
 Romein, J. 375, 377, 399
 Romney, D. 173, 177, 202
 Rorer, L. G. 231, 299
 Roscoe, A. 278, 300
 Rosenbaum, L. L. 169, 216
 Rosenbaum, W. B. 216
 Rosenblum, A. L. 286, 298
 Rosenblum, L. A. 13, 68
 Rosenmeier, H. P. 334, 359
 Rosenthal, H. 66
 Rosenthal, R. 16, 72, 120, 132
 Roslow, S. 244, 247, 257, 260, 266, 299
 Rosnow, R. L. 16, 71, 120, 132
 Ross, B. M. 216
 Ross, J. 185, 216
 Ross, M. 317, 320
 Rothenberg, A. 297
 Rothman, A. I. 216

 Rowe, P. M. 329, 359
 Rubin, M. 140, 152
 Robinson, R. 66
 Rudinger, G. 13, 42, 72, 104, 132
 Rugg, D. 254f, 257f, 261, 266, 271, 299
 Rugg, H. 104, 132
 Ruggels, W. L. 202
 Ruppel, M. 364, 392, 399
 Ruppenthal, G. C. 72
 Russell, W. A. 209
 Rydell, S. T. 178, 216
 Rytten, B. 212

 Saari, B. B. 71
 Sackett, G. P. 18, 19, 72
 Sagara, M. 216
 Salapatek, P. H. 365, 396
 Salber, W. 325, 359
 Sappenfield, B. R. 146, 152
 Sarason, I. G. 308, 319
 Schäfer, B. 171, 184, 188, 193, 194, 199, 204, 216
 Schapp, W. 92, 132
 Scharnweber 374, 399
 Scheele, B. 317, 320
 Schefflen, A. E. 100, 132
 Scheier, I. H. 225, 299
 Scheirer, C. J. 107, 131
 Schenck, E. A. 17, 71
 Scherer, K. R. 2, 72, 306, 320
 Scheuch, E. 322, 359
 Seheuch, E. K. 223, 228, 241, 245, 251, 261, 286, 299, 300, 302, 305, 320
 Schiavo, R. S. 320
 Schick, A. 191, 192, 216
 Schlosberg, H. 164, 216
 Schludermann, E. 216
 Schludermann, S. 178, 216

- Schmidt, H. D. 325,
334, 347, 359, 360, 367,
399
- Schmitt, N. 56, 58, 60,
72
- Schneider, J. 229, 233,
264, 300
- Schneider-Düker, M.
264, 300
- Schön, G.-H. 152
- Schoenberg, R. 57, 63
- Schönflug, W. 198, 216
- Schoggen, P. 20, 72,
100, 124
- Schonfeld, W. A. 346,
359
- Schraml, W. 325, 333,
334, 359, 370, 399
- Schramm, W. 168, 210
- Schreiber, K. 241, 300
- Schriesheim, C. 229, 300
- Schriesheim, J. 300
- Schucany, W. R. 36, 72
- Schümer, R. 105, 126,
132
- Schuh, A. J. 217
- Schuller, A. 334, 359
- Schulter, G. 178, 191,
192, 212
- Schulz, R. 296
- Schulz, W. 356
- Schulz v. Thun, F. 135,
151
- Schumann, H. 312, 320
- Schutz, W. C. 29, 72
- Schwartz, C. G. 5, 72
- Schwartz, M. S. 5, 72
- Schwartz, R. D. 74, 133
- Schwarzer, R. 308, 320
- Schwenkmezger, P. 295
- Schyberger, B. W. 245,
300
- Scott, C. 307, 320
- Scott, R. A. 60
- Scott, W. A. 16, 30, 73
- Scupin, E. 365, 399
- Scupin, G. 365, 399
- Sears, R. R. 368, 399
- Sechrest, L. 74, 133
- Seeman, W. 140, 151
- Seidman, E. 56, 58, 65
- Selg, H. 2, 71
- Seligman, E. 373, 399
- Selvage, R. 37, 73
- Semmel, M. 1. 26, 65
- Settle, R. B. 35, 74
- Seyfried, B. A. 295
- Shapiro, M. B. 148, 152
- Sharma, S. N. 280, 300
- Shaw, M. E. 179, 213
- Shaw, R. 378
- Sheatsley, P. B. 241, 300
- Shell, S. A. 217
- Shepard, R. N. 95f, 131,
132
- Shepherd, I. L. 140, 152
- Sherif, M. 188, 208
- Sherman, R. E. 392, 397
- Sherry, P. 140, 152
- Sheth, J. 278, 300
- Shikiar, R. 168, 172, 217
- Shinn, M. W. 365, 399
- Shirk, E. J. 203
- Shlien, J. M. 140, 152
- Shontz, F. C. 140, 152
- Shrout, P. E. 38, 65
- Sicoly, F. 317, 320
- Sieber, M. 287, 289, 292,
300, 307, 19
- Sieveking, N. A. 296
- Simkins, L. 119, 132
- Simon, A. 13, 73
- Simon, H. A. 91, 14,
116, 127
- Simons, G. 2, 73
- Simons, J. L. 5, 6, 70
- Simonton, D. K. 386,
399f
- Simpson, R. H. 229, 00
- Sines, J. O. 217
- Singer, R. D. 217
- Singh, Y. P. 280, 00
- Singleton, W. T. 73
- Sixtl, F. 231, 300
- Skinner, B. F. 89f, 93,
100, 132
- Slobin, D. I. 205
- Smith, E. R. 320
- Smith, F. V. 207
- Smith, G. 220
- Smith, R. G. 191, 217
- Snider, J. G. 154, 217
- Snyder, F. 171, 172,
175, 179, 217
- Snyder, F. W. 217
- Snyder, W. U. 152
- Sörbom, D. 43, 56, 67,
68, 304, 319
- Solarz, A. K. 217
- Solle, R. 210
- Somers, R. H. 73
- Sommer, R. 196, 217
- Sorembe, V. 41, 73
- Soudijn, K. A. 334, 359
- Spearman, C. 34, 36, 73
- Spiegel, B. 360
- Spinner, B. 98, 124
- Spitzer, R. L. 65
- Spoerer, E. 284, 400
- Spranger, E. 363, 400
- Spriegel, W. R. 324, 360
- Springbett, B. M. 217
- Staats, A. W. 166, 205,
218
- Staats, C. K. 218
- Stäcker, K. H. 250, 295
- Stahlberg, G. 213
- Stanley, J. C. 16, 57, 58,
62, 73, 107, 133, 380,
389, 395
- Starr, D.J. 105, 133
- Steinkamp, S. W. 329,
360
- Steller, M. 144, 145, 152
- Stephenson, W. 135,
140, 142, 153
- Stern, C. 365, 400
- Stern, W. 85, 133, 323,
324, 356, 360, 365, 400
- Steward, C. J. 241, 300
- Steward, R. B. 73
- Steward, T. R. 133
- Stewart, R. A. 36, 73

- Stollberger, R. 242, 246, 275, 300
 Stover, D. O. 204
 Strahan, R. 229, 231, 300
 Straka, J. 146, 151
 Stricker, G. 178, 218
 Stricker, L. J. 105, 133
 Stroebe, W. 180, 204
 Strong, E. R. 289, 300
 Stroschein, F. R. 241, 244ff, 250, 254, 256, 261, 266, 268f, 274, 285, 301
 Stroud, T. W. F. 48, 73
 Subotnik, L. 146, 153
 Suchman, E. A. 266, 301
 Suci, G. 168, 188, 215
 Suci, G. J. 154, 209, 218, 298
 Sudman, S. 229, 247f, 265, 283f, 292, 294, 301, 307, 320
 Süllwold, F. 228, 301
 Suk, J. M. 310, 320
 Summers, G. F. 73
 Susman, E. J. 26, 73
 Sutcliffe, J. P. 48, 73
 Swaminathan, H. 387, 400
 Swan, J. E. 284, 294
 Swets, J. A. 28, 65, 107, 133

 Tack, W. H. 400
 Taft, R. 329, 360
 Tagiuri, R. 104, 125
 Taietz, P. 283, 301
 Taine, H. M. 365, 400
 Tajfel, H. 181, 218
 Takahashi, S. 218
 Tamulonis, V. 292
 Tanaka, Y. 162, 170, 172, 215, 218
 Tannenbaum, P. 215
 Tannenbaum, P. H. 154, 199, 204, 218, 298
 Tansill, R. 212

 Taplin, P. S. 118f, 133
 Tatsuoka, M. M. 292
 Taubert, H. 296
 Taylor, C. L. 184, 218
 Taylor, D. M. 146, 153, 206, 218
 Taylor, H. F. 219
 Taylor, R. E. 204
 Taylor, W. L. 36, 73
 Terborg, J. R. 260, 301
 Terwilliger, R. F. 185, 202, 219
 Tesser, A. 50, 74
 Thackray, R. I. 333, 359
 Tholey, V. 229, 301
 Thomae, H. 2, 13, 74, 88, 133, 322, 323, 331, 333, 343, 348, 353, 360, 375, 397, 400
 Thomas, W. I. 371, 378, 400
 Thompson, C. 212
 Thoms, K. 322, 360
 Thorndike, E. L. 104, 133
 Thorne, F. C. 334, 360
 Thornton, G. C. 220
 Thumin, F. J. 201
 Tiedemann, D. 365, 400
 Timaeus, E. 207
 Tinsley, H. E. A. 26, 40, 74
 Titchener, E. B. 81, 133
 Titscher, S. 241f, 246, 251, 255, 257ff, 262ff, 267, 269f, 274f, 286, 297
 Tittle, C. R. 273, 301
 Tobacyk, J. J. 140, 153
 Tolman, E. C. 110, 131
 Tränkle, U. 223, 282, 285, 301
 Trankell, A. 324, 350, 360
 Triandis, H. C. 168, 192, 199, 219
 Triebe, J. K. 354, 355, 360

 Triebe, K. 320
 Trippi, R. R. 35, 74
 True, J. E. 284, 296
 Trumbo, D. 341, 353, 360
 Trush, R. S. 146, 153
 Tucker, L. R. 56, 58, 74
 Turner, C. 229, 301
 Turner, R. H. 140, 153
 Turvey, M. T. 184, 219
 Tversky, A. 126
 Tzeng, O. C. S. 165, 171, 172, 176, 183, 199, 219

 Ulich, E. 354, 355, 360
 Ulrich, L. 341, 353, 360
 Underwood, W. L. 220
 Undeutsch, U. 325, 326, 334, 360
 Utz, H. 295

 Van Atta, R. E. 140, 153
 Van der Kamp, L. J. T. 109, 129, 133
 Vanderlippe, R. H. 140, 153
 Van Meter, D. 107, 133
 Vaught, G. M. 140, 153
 Vegelius, J. 33, 67, 74
 Verinis, J. S. 216
 Verner, H. W. 292
 Vernon, Ph. E. 338, 360
 Vidal, J. J. 183, 197, 219
 Villamin, A. C. 206
 Vincent, R. A. 144, 153
 Vitale, J. 62, 74
 Vogel, B. 392, 399
 Voyce, C. D. 180, 219

 Wackerly, D. D. 31, 32, 74
 Wagner, R. 341, 361
 Wahl, D. 316, 320
 Walker, B. A. 219

- Walker, R. N. 153
 Wall, D. D. 179, 209
 Wallbott, H. G. 26, 32, 37, 61
 Walther, E. H. 352, 361
 Ware, E. E. 162, 170, 211, 215, 220
 Warr, P. B. 199, 220
 Warren, J. T. 66
 Washington, W. N. 178, 220
 Waskow, J. E. 147, 148, 153
 Watkins, M. W. 31, 70, 74
 Watson 89
 Wattawa, S. 290, 293
 Webb, E. J. 74, 120, 133
 Weick, K. E. 2, 20, 74
 Weigel, R. G. 220
 Weigel, V. M. 220
 Weimer, J. 208
 Weinreich, U. 220
 Weiss, D.J. 24, 26, 40, 74
 Weksel, W. 220
 Wellek, A. 326, 339, 352, 353, 355, 361
 Wells, F. L. 104, 133
 Wells, H. G. 378
 Wells, W. D. 220
 Welzel, U. 328, 350, 361
 Werner, J. 37, 38, 74
 Wertheimer, M. 73, 93, 133
 Werts, C. E. 42, 43, 48, 56, 71, 74, 75, 110, 133
 Whaley, F. 71
 Wheaton, B. 43, 75
 Whelan, P. 75
 White, G. 122, 128
 White, P. 90, 133
 White, R. W. 400
 Whiting, B. B. 100, 133
 Whiting, J. W. 100, 133
 Whiting, J. W. M. 17, 75
 Whitney, D. R. 290, 301
 Whyte, W. F. 5, 75, 332, 361
 Wichmann, U. 295
 Wickens, D. D. 184, 220
 Wieczorek, R. 205
 Wieken, K. 241, 276, 285, 286f, 301, 305, 320
 Wieken-Mayser, M. 297
 Wiendieck, G. 310, 319
 Wieser, I. 400
 Wiggins, N. 170, 171, 172, 175, 176, 179, 208, 217, 220
 Wilbur, P. H. 144, 153
 Wilcox, R. C. 220
 Wilder, D. 113, 133
 Wildman, R. C. 284, 301
 Wildman, R. W. 220
 Wiley, D. E. 43, 75
 Wiley, J. A. 43, 75
 Wiley, L. 329, 361
 Wilk, G. 241, 301
 Williams, J. E. 220
 Williams, W. S. 170, 220, 221
 Willick, D. H. 271, 278, 301
 Wilson, R. N. 368, 400
 Wilson, T. D. 79, 131, 316, 320
 Wilson, T. P. 36, 75
 Winer, B. J. 38, 75
 Winograd, E. 221
 Wittenborn, J. R. 135, 147, 148, 153
 Wittrock, M. C. 208
 Wohlfart, E. 361
 Wolf, G. 66
 Wolf, M. M. 127, 394, 399
 Wolfe, L. A. 334, 361
 Wolfenstein, E. v. 370, 400
 Wolfson, A. D. 297
 Wolins, L. 56, 62, 68
 Wonnacott, E. J. 206
 Woodward, J. A. 67
 Woog, P. C. 146, 153
 Wottawa, H. 226f, 262, 266, 268, 287, 301
 Wright, H. F. 2, 18, 20, 75, 377, 394
 Wright, O. R. 353, 355, 361
 Wright, P. 263f, 281, 301
 Wulfeck, H. 299
 Wundt, W. 80, 82f, 83, 134
 Wyckoff, D. 293
 Wylie, R. C. 138, 144, 145, 146, 153
 Wynd, W. 293
 Yamamoto, K. 216
 Yates, F. 16, 75
 Young, D. D. 221
 Younger, M. S. 43, 44, 70
 Zander, A. F. 2, 66
 Zaniecki, F. 371, 378, 400
 Zavalloni, M. 180, 221
 Zax, M. 178, 211, 218, 221
 Zehnpfennig, H. 300
 Ziller, R. C. 211
 Zippel, B. 221